

501.43680X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: K. SHIGA, et al

Serial No.: 10/801,720

Filing Date: March 17, 2004

For: PATH CONTROL METHOD

LETTER CLAIMING RIGHT OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

June 8, 2004

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, applicants hereby claim
the right of priority based on:

**Japanese Application No. 2004-022463
Filed: January 30, 2004**

A Certified copy of said application document is attached hereto.

Acknowledgement thereof is respectfully requested.

Respectfully submitted,

Carl I. Brundidge
Registration No. 29,621
ANTONELLI, TERRY, STOUT & KRAUS, LLP

CIB/jdc
Enclosures
703/312-6600

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 4 年 1 月 3 0 日
Date of Application:

出 願 番 号 特 願 2 0 0 4 - 0 2 2 4 6 3
Application Number:
[ST. 10/C] : [J P 2 0 0 4 - 0 2 2 4 6 3]

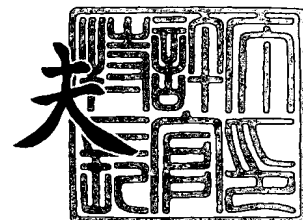
出 願 人 株式会社日立製作所
Applicant(s):



特許庁長官
Commissioner,
Japan Patent Office

2 0 0 4 年 3 月 2 日

今 井 康 夫



出証番号 出証特 2 0 0 4 - 3 0 1 5 6 0 9

【書類名】 特許願
【整理番号】 K03018001A
【あて先】 特許庁長官殿
【国際特許分類】 G06F 13/14
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 志賀 賢太
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 熊谷 敦也
【発明者】
 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所
 システム開発研究所内
 【氏名】 藤原 啓成
【特許出願人】
 【識別番号】 000005108
 【氏名又は名称】 株式会社 日立製作所
【代理人】
 【識別番号】 100075096
 【弁理士】
 【氏名又は名称】 作田 康夫
【選任した代理人】
 【識別番号】 100100310
 【弁理士】
 【氏名又は名称】 井上 学
【手数料の表示】
 【予納台帳番号】 013088
 【納付金額】 21,000円
【提出物件の目録】
 【物件名】 特許請求の範囲 1
 【物件名】 明細書 1
 【物件名】 図面 1
 【物件名】 要約書 1

【書類名】 特許請求の範囲**【請求項 1】**

第一の装置と、
第二の装置と、
前記第一の装置と前記第二の装置とを接続する複数のパスと、
前記第一の装置と接続される第三の装置とを有し、
前記第一の装置は、所定の割合で前記複数のパスを使用して前記第二の装置へデータを転送し、
前記第三の装置は、前記複数のパスの輻輳を検出して前記第一の装置へ通知し、
前記第一の装置は、前記通知に基づいて、前記所定のパス間の割合を変更して前記複数のパスを使用して前記第二の装置へデータを転送することを特徴とするシステム。

【請求項 2】

前記第一の装置及び前記第二の装置はストレージ装置であることを特徴とする請求項 1 記載のシステム。

【請求項 3】

前記第一の装置は計算機であり、前記第二の装置はストレージ装置であることを特徴とする請求項 1 記載のシステム。

【請求項 4】

前記複数のパスの各々は前記第一の装置と前記第二の装置とを接続するためのネットワーク装置を有し、
前記第三の装置は、前記ネットワーク装置とネットワークを介して接続されており、
前記第三の装置は、前記ネットワーク装置から前記ネットワークを介して前記ネットワーク装置での輻輳の発生のお知らせを受け取ることを特徴とする請求項 2 及び 3 のうちいずれか一つに記載のシステム。

【請求項 5】

前記通知とは、SNMP Trap に基づく通知であることを特徴とする請求項 4 記載のシステム。

【請求項 6】

前記複数のパスの各々は前記第一の装置と前記第二の装置とを接続するためのネットワーク装置を有し、
前記第三の装置は、前記ネットワーク装置とネットワークを介して接続されており、
前記第三の装置は、前記ネットワーク装置から前記ネットワークを介して前記ネットワーク装置での廃棄パケットに関する情報を受け取り、前記廃棄パケットの情報に基づいて前記複数のパスの輻輳を判断することを特徴とする請求項 2 及び 3 のうちいずれか一つに記載のシステム。

【請求項 7】

前記第三の装置は、前記ネットワーク装置から受信する廃棄パケット数が、以前受信した廃棄パケット数を上回る場合に前記ネットワーク装置を有する前記複数のパスに輻輳が発生したと判断することを特徴とする請求項 6 記載のシステム。

【請求項 8】

前記第三の装置は、前記所定の割合及び前記所定の割合の変更率に関する情報を有し、前記複数のパスの輻輳を検出した場合、変更後の前記所定のパス間の割合を前記変更率に基づいて計算して、前記第一の装置に前記変更後の所定の割合の情報を送信し、

前記第一の装置は、受信した前記変更後の所定のパス間の割合に基づいて、前記複数のパスを用いて前記第二の装置へデータを転送することを特徴とする請求項 1 記載のシステム。

【請求項 9】

前記第三の装置は、前記複数のパスの輻輳からの回復を検知し、前記所定のパス間の割合の情報を前記第一の装置に送信し、

前記第一の装置は、受信した前記所定のパス間の割合に基づいて、前記複数のパスを用

いて前記第二の装置へデータを転送することを特徴とする請求項 8 記載のシステム。

【請求項 10】

第一の装置と、
第二の装置と、
前記第一の装置と前記第二の装置とを接続する複数のパスとを有し、
前記第一の装置は、所定の割合で前記複数のパスを使用して前記第二の装置へデータを転送し、
前記第一の装置は、前記複数のパスの輻輳を検出し、
前記第一の装置は、前記複数のパスの輻輳の検出に応じて、前記所定のパス間の割合を変更して前記複数のパスを使用して前記第二の装置へデータを転送することを特徴とするシステム。

【請求項 11】

前記第一の装置及び前記第二の装置はストレージ装置であることを特徴とする請求項 10 記載のシステム。

【請求項 12】

前記第一の装置は計算機、前記第二の装置はストレージ装置であることを特徴とする請求項 10 記載のシステム。

【請求項 13】

前記複数のパスの各々は前記第一の装置と前記第二の装置とを接続するためのネットワーク装置を有し、
前記第一の装置は、前記ネットワーク装置から前記複数のパスを介して前記ネットワーク装置での輻輳の発生のお知らせを受け取ることを特徴とする請求項 11 及び 12 のうちいずれか一つに記載のシステム。

【請求項 14】

前記通知とは、ECNに基づくフラグであることを特徴とする請求項 13 記載のシステム。

【請求項 15】

前記第一の装置は、前記第二の装置から所定の期間応答が帰ってこない場合に前記複数のパスに輻輳が発生したと判断することを特徴とする請求項 11 及び 12 のうちいずれか一つに記載のシステム。

【請求項 16】

前記第一の装置は、前記第二の装置へ送信したデータの受信確認応答を重複して受信した場合に、前記複数のパスに輻輳が発生したと判断することを特徴とする請求項 11 及び 12 のうちいずれか一つに記載のシステム。

【請求項 17】

前記第一の装置は、前記所定の割合及び前記所定の割合の変更率に関する情報を有し、前記複数のパスの輻輳を検出した場合、変更後の前記所定のパス間の割合を前記変更率に基づいて計算し、前記変更後の所定のパス間の割合に基づいて、前記複数のパスを用いて前記第二の装置へデータを転送することを特徴とする請求項 10 記載のシステム。

【請求項 18】

前記第一の装置は、輻輳発生後、輻輳が発生した前記複数のパスに送信できるデータサイズがあらかじめ設定された値を超過したときに、輻輳が回復したと判断することを特徴とする請求項 17 記載のシステム。

【請求項 19】

制御部と、
前記制御部に接続されるディスク装置と、
ネットワークに接続されるインターフェースとを有し、
前記インターフェースは、前記ネットワーク内の複数のパスで他の装置と接続され、
前記制御部は所定のパス間の割合で前記複数のパスを用いて前記他の装置に前記ディスク装置に格納されたデータをパケットにして送信し、

前記制御部は、前記他の装置に送信したパケットに対する受領報告が一定期間受信されない場合に、前記複数のパスに輻輳が発生したと判断し、

前記制御部は、前記輻輳の発生に応じて、前記所定のパス間の割合を変更し、前記変更されたパス間の割合で前記他の装置へのパケット転送を行うことを特徴とするストレージ装置。

【書類名】明細書

【発明の名称】パス制御方法

【技術分野】

【0001】

本発明は、ネットワークを介して接続されたストレージ装置と情報処理装置を有するシステムに関する。

【背景技術】

【0002】

近年、計算機（以下「ホスト」ともいう）にストレージ装置を直結する形態の代わりに、一つ又は複数のストレージ装置を、ネットワークを介して複数のホストと接続する形態が盛んに用いられるようになってきている。このネットワークを用いたストレージ装置の接続形態のことをストレージエリアネットワーク（以下「SAN」）という。これまでSANはファイバチャネル（以下「FC」）技術を用いて構築されていた。以下、FCベースのSANをFC-SANと称する。

【0003】

FC-SANを構築する場合、システム全体の可用性を向上させるために、ホストとストレージ装置との間は複数の物理的な通信経路（以下「パス」ともいう）で接続される。ホストは、ストレージ装置へ送信するコマンドやデータを各々のパスを使用して送信することで、パスに掛かる負荷を分散させる。複数のパスの使用割合を決めるアルゴリズムとして、送信されるコマンドやデータを各々のパスに異なる割合（例えば送信されるコマンド数の6割：4割）で割り振る重み付けラウンドロビンが挙げられる。またホストは、FC-SANを構成するスイッチの故障やネットワークの切断等の原因により1つのパスが使用不能になったことを検出した場合、残りのパスを使用してストレージ装置との通信を継続する。

【0004】

上記のようなパス間の負荷分散は、ホストとストレージ装置との間の通信（以下「ホストストレージ装置間通信」）だけではなく、ストレージ装置と他のストレージ装置との間の通信（以下「ストレージ装置間通信」）でも実施されている。ストレージ装置間通信の例としてリモートコピーが挙げられる。リモートコピーとは、ストレージ装置に格納されるデータを、地理的に離れた拠点に設置された他のストレージ装置へ転送してデータを複製保存する技術である。

【0005】

上述したホストストレージ装置間通信におけるパス制御技術の1つが特許文献1に開示されている。

【0006】

一方、FCよりも導入コストが低いインターネットプロトコル（以下「IP」）を用いたネットワーク（以下「IPネットワーク」）を用いて構築されたSANであるIP-SANが注目を集めている。ホストがIP-SANを介してストレージ装置と通信する場合、SCSIプロトコルをTCP/IPでカプセル化したプロトコルであるiSCSIが主に用いられる。

【0007】

【特許文献1】特開2000-330924号公報

【発明の開示】

【発明が解決しようとする課題】

【0008】

IP-SANにおいてホストストレージ装置間通信又はストレージ装置間通信をする場合、特許文献1に開示されたパス制御技術では以下の問題が発生する。

上述したように、特許文献1ではネットワークとしてFCを使用する。FCでは、通信におけるフロー制御にバッファクレジットと呼ばれる技術が用いられる。バッファクレジットでは、ホストやストレージ装置などのネットワークの末端（以下「エンドポイント」

）と、そのエンドポイントが接続されているスイッチとが、受信したパケットを一時的に保存するためのバッファの空きサイズの情報をお互いに交換する。同様に、隣接したスイッチ同士もバッファの空きサイズの情報をお互いに交換する。これによりネットワーク間でバッファの空きサイズ以上のパケットを送信しない様に各装置がデータを転送する。このため、FCでは基本的にネットワーク内での輻輳が発生しない。したがって、特許文献1ではそもそも輻輳に基づいてパスを制御することは全く考慮されていない。

【0009】

一方、TCP/IPでは、フロー制御にスライディングウィンドウと呼ばれる技術が用いられる。スライディングウィンドウでは、エンドポイント同士がバッファの空きサイズの情報をお互いに交換する。ここで、エンドポイントはネットワークを構成するスイッチのバッファの空きサイズを考慮せずにパケットを送信する。したがって、ネットワークに送出されたパケットがスイッチのバッファの許容量を超過した場合、輻輳が発生する。輻輳が発生すると、スイッチは受信したパケットを廃棄してしまう。更にエンドポイントは、タイムアウト等の手段で輻輳の発生を検出すると、輻輳を回復させるためにパケットの送信レートを大幅に下げる。この仕組みは、RFC 2581により規定されている。

【0010】

従って、輻輳が発生しないFCを前提とした従来のパス制御技術をIP-SANに適用すると、TCP/IPの仕組みそのままに、輻輳が発生したパスの転送レートが大幅に下がってしまう。しかし、特許文献1等で開示された従来のパス制御技術では輻輳を考慮しないので、各エンドポイントはその送信レートが低下したパスを同じ使用率で使用し続ける。したがってシステム全体のパケット送信のスループットが大幅に下がってしまう。特にストレージ装置へのデータの入出力の際にはスループットが低下することは大きな問題（書き込み応答の遅延等）となる。

【課題を解決するための手段】

【0011】

そこで、本発明の一実施形態として、第一の装置と、第二の装置と、第一の装置と前記第二の装置とを接続する複数のパスと、第一の装置と接続される第三の装置とを有するシステムで、第一の装置は、所定の割合で複数のパスを使用して第二の装置へデータを転送し、第三の装置は、複数のパスの輻輳を検出して第一の装置へ通知し、第一の装置は、通知に基づいて、所定のパス間の割合を変更して複数のパスを使用して第二の装置へデータを転送する構成とする。

【0012】

尚、本発明の実施態様として、第一の装置は計算機でもストレージ装置でも良い。又、第三の装置は第一の装置に含まれていても良い。

【0013】

更に、本発明の実施態様として、第三の装置で所定の割合及び所定の割合の変更率に関する情報を有し、複数のパスの輻輳を検出した場合、変更後の前記所定のパス間の割合を変更率に基づいて計算し、変更後の所定のパス間の割合の情報を第一の装置に送信し、第一の装置は、変更後の所定のパス間の割合に基づいて、複数のパスを用いて第二の装置へデータを転送する構成でも良い。

【0014】

更に、本発明の実施態様として、複数のパスは第一の装置と第二の装置を接続するネットワーク装置を有し、第三の装置はネットワーク装置から輻輳の発生や輻輳からの回復についての情報を得る構成も考えられる。その他の実施態様については、以下の記載で明らかにされる。

【発明の効果】

【0015】

本発明により、システム全体のパケット送信スループット、具体的にはホストストレージ装置間又はストレージ装置間のパケット転送スループットの低下を最小限に抑えることができる。

【発明を実施するための最良の形態】**【0016】**

以下、本発明の実施形態について図面を用いて説明する。以下の図中、同一の部分には同一の符号を付加する。ただし本発明が実施形態に制限されることは無く、本発明の思想に合致するあらゆる応用例が本発明に該当する。また、特に限定しない限り、各構成要素は複数でも単数でも構わない。

【実施例1】**【0017】**

第一の実施形態は、ストレージ装置間で実行される iSCSI を用いたリモートコピーに本発明を適用したシステムに関する。第一の実施形態では、一方のストレージ装置（以下「マスタストレージ装置」）が、他方のストレージ装置（以下「リモートストレージ装置」）と1つ以上のTCPコネクションから成る1つの iSCSI セッションを確立してリモートコピーを実行する。それぞれのTCPコネクションを用いて送信されるパケットが通る経路が1つのパスに相当する。

【0018】

図1は、第一の実施形態のシステムの構成例を示した図である。システムは、マスタサイト11、リモートサイト12およびこれらのサイト間を接続する1つ以上のWAN（Wide Area Network）8を有する。なお「サイト」とは装置が設置される場所や装置群を指し、一つのビル等を指す。また本実施形態では、マスタサイト11とリモートサイト12とはある程度距離が離れているもの（例えば東京と大阪等）とする。

【0019】

マスタサイト11は、マスタストレージ装置1、マスタストレージ装置1を管理するストレージ管理装置2、ホスト4、端末6、ホスト4、マスタストレージ装置1およびストレージ管理装置2を接続するIPネットワークであるIP-SAN7並びに端末6とホスト4とを接続するIPネットワークであるLAN9を有する。また、マスタストレージ装置1、ストレージ管理装置2およびホスト4とIP-SAN7とは、UTP（Unshielded Twisted Pair）ケーブルや光ファイバケーブル等の通信線10で接続されている。同様に、端末6およびホスト4とLAN9とは、通信線10で接続されている。

【0020】

なお、端末6等の装置と、IP-SAN7やLAN9等のIPネットワークとを、無線通信技術を使用して接続する場合、通信線10は不要である。また、本実施形態ではIP-SAN7とLAN9とが各々別である例を示すが、IP-SAN7がLAN9を兼ねる、すなわち全ての装置がIP-SAN7に接続される構成としても良い。この場合にはシステム設置のコストは抑えられる。しかしストレージ装置間通信用のパケットと端末間での通信に使用されるパケットが一つのネットワークに混在してネットワークが混雑するという問題がある。これを解消するためには本実施形態の構成が好ましい。

【0021】

リモートサイト12は、リモートストレージ装置5、ホスト4、端末6、ホスト4とリモートストレージ装置5とを接続するIP-SAN7、端末6とホスト4とを接続するLAN9を有する。リモートストレージ装置5およびホスト4とIP-SAN7とは、通信線10で接続されている。同様に、端末6およびホスト4とLAN9とは、通信線10で接続されている。リモートサイト12でも、ネットワークが一つである構成も考えられる。

【0022】

ホスト4および端末6は一般的な計算機であり、CPU、主メモリおよび入出力装置等を有する。また、ホスト4および端末6は、他の装置と通信線10を介して接続するインタフェースであるネットワークインタフェース（以下「NIF」）を有する。

WAN8は、受信したパケットを転送する1つ以上のネットワーク装置3を有する。

【0023】

図2(a)は、マスタストレージ装置1の構成例を示した図である。マスタストレージ装置1は、単体の記憶装置あるいは複数の記憶装置を有する記憶装置システムである。なお、記憶装置は、ハードディスクドライブやDVDと言った、不揮発性の記憶媒体を用いた装置が含まれる。また、記憶装置システムでは、RAID構成が採用されていてもよい。マスタストレージ装置1は、記憶装置(以下「ディスク装置」)103、ディスク装置103に対するデータの書き込みや読み出しを制御する制御装置107および制御装置107とディスク装置103とを接続する通信線106を有する。

【0024】

制御装置107は、揮発性のメモリ(以下「主メモリ」)101、バスなどの通信線102、中央演算装置(以下「CPU」)104、制御装置107と通信線106とを接続するインタフェースであるIOインタフェース(以下「IO IF」)105および制御装置107と通信線10とを接続するNIF108を有する。

【0025】

主メモリ101には、ディスク装置103から読み出されたデータ又はホスト等から受信したデータを記憶するキャッシュ領域110、リモートコピーを実行する際にCPU104で実行されるリモートコピー実行プログラム111、ストレージ管理装置2から割合変更要求を受信した際にCPU104で実行される割合変更要求受信プログラム112が格納される。

【0026】

図2(b)は、ストレージ管理装置2の構成例を示した図である。ストレージ管理装置2は、主メモリ201、通信線202、ディスク装置203、CPU204、表示装置などの出力装置(以下「ディスプレイ」)205、マウスなどのポインティング装置206、キーボードなどの文字入力装置207およびNIF208を有する計算機である。

【0027】

主メモリ201には、ネットワーク装置3から輻輳発生通知や輻輳回復通知を受信した際にCPU204で実行される輻輳受信プログラム210、バス間の負荷分散の割合を計算する際にCPU204で実行される割合計算プログラム211、マスタストレージ装置1等へ割合変更要求を送信する際にCPU204で実行される割合変更プログラム212およびシステム管理者等にグラフィカルなユーザインターフェースを提供する際にCPU204で実行されるGUI制御プログラム213が格納される。

【0028】

なお、ここで「バス間の負荷分散の割合」(以下単に「バス間の割合」又は「バスの割合」ということもある)とは、ホストストレージ装置間通信やストレージ装置間通信で利用されるバス群でどのようにコマンドやデータの転送量(あるいは数)を分担するかを示す数値である。具体的には、ある一つの処理において3本のバスを使用する場合、「バス間の負荷分散の割合」とは、当該処理に含まれる転送パケット数のうちの3割を第一のバスで、3割を第二のバスで、4割を第三のバスで送信する場合の「3割、3割、4割」の組又はその組の個々の数値を示す。

【0029】

また、ディスク装置203には、WAN8を構成するネットワーク装置3の情報を記憶するネットワーク管理テーブル220、マスタストレージ装置1がリモートコピーの個々のタスクを実行するのに必要な情報を記憶するタスク管理テーブル221およびマスタストレージ装置1とリモートストレージ装置5との間で使用されるバスの情報を記憶するバス管理テーブル222が格納される。ここで、タスクとは、マスタストレージ装置1が有する論理ユニット(Logical Unit: LU)から、リモートストレージ装置2が有する論理ユニットへのリモートコピーに関わるデータ転送処理を指す。なお、LUとはディスク装置が有する物理的な記憶領域から構成される論理的な記憶領域である。LUは一つのディスク装置が有する記憶領域から構成されてもよく、複数のディスク装置の個々の記憶領域の集合体として定義されても良い。

【0030】

なお、第一の実施形態では、ストレージ管理装置 2 にて障害が発生した場合でも各テーブルが記憶する情報が喪失されないようにディスク装置 203 に各テーブルが格納されることとしたが、各テーブルの情報が主メモリ 201 に格納されても良い。

【0031】

なお、リモートストレージ装置 5 は、主メモリ 101 に割合変更要求受信プログラム 112 が格納されない点を除き、マスタストレージ装置 1 と同様の構成を有する。

【0032】

図 3 は、ネットワーク装置 3 の構成例を示した図である。ネットワーク装置 3 は、主メモリ 301、通信線 302、CPU 304、NIF 308 および NIF 308 が受信したパケットを別の NIF 308 を介して他の装置へ送信するパケット転送装置 309 を有する。主メモリ 301 には、転送待ちパケットを一時的に格納するバッファ領域 310 およびバッファ領域 310 の容量が不足した時にストレージ管理装置 2 へ輻輳発生通知を送信する際やバッファ領域 310 の容量が空いた時にストレージ管理装置 2 へ輻輳回復通知を送信する際に CPU 304 で実行される輻輳状態通知プログラム 311 が格納される。

【0033】

なお、上述のプログラムは、あらかじめ、または可搬型記録媒体からの読み込み、または他の計算機からのネットワーク経由のダウンロードにより、ディスク装置やメモリに格納される。これらのプログラムは、必要に応じて主メモリに転送され、CPU で実行される。

【0034】

次に、ストレージ管理装置 2 のディスク装置 203 に格納される各種テーブルのデータ構造について説明する。

【0035】

ネットワーク管理テーブル 220、タスク管理テーブル 221 およびパス管理テーブル 222 は配列構造を成し、1 つ以上のレコードを格納可能である。ただし、データ構造が配列構造に限定されることはない。

【0036】

図 4 (a) は、ネットワーク管理テーブル 220 のデータ構造例を示す図である。ネットワーク管理テーブル 220 は、マスタストレージ装置 1 とリモートストレージ装置 5 との通信に関わるネットワーク装置 3 の個数分のレコードを有する。ネットワーク管理テーブル 220 の各レコードは、対応するネットワーク装置 3 が属するネットワークを識別する為の識別子であるネットワーク ID が登録されるエン트리 2201、ネットワーク装置 3 を識別するための識別子である装置 ID が登録されるエン트리 2202 およびネットワーク装置 3 が送信する輻輳発生通知や輻輳回復通知のソース IP アドレスが登録されるエン트리 2203 を有する。

【0037】

図 4 (b) は、タスク管理テーブル 221 のデータ構造例を示す図である。タスク管理テーブル 221 は、リモートコピーのタスクごとに 1 つのレコードを有する。タスク管理テーブル 221 の各レコードは、個々のリモートコピータスクを識別するための識別子であるタスク ID が登録されるエン트리 2211、iSCSI イニシエータとして動作するマスタストレージ装置 1 が有する iSCSI 名であるイニシエータ iSCSI 名が登録されるエン트리 2212、コピー対象となる論理ユニットを識別するための識別子であるイニシエータ LU 番号 (LUN) が登録されるエン트리 2213、iSCSI ターゲットとして動作するリモートストレージ装置 5 が有する iSCSI 名であるターゲット iSCSI 名が登録されるエン트리 2214、コピー先の LU を識別するためのターゲット LUN が登録されるエン트리 2215、輻輳が発生した時にパス間の負荷分散の割合を変更するか否かを示す割合変更フラグが登録されるエン트리 2216 および輻輳が発生したパスの負荷分散の割合の変更率を示す割合変更率が登録されるエン트리 2217 を有する。

【0038】

第一の実施形態では、エン트리 2216 の「0」は輻輳時に割合変更を行わないことを

、「1」は輻輳時に割合変更を行うことをそれぞれ表すものとする。また、第一の実施形態では、エントリ 2217 には、割合変更率を百分率で表した値が登録されるものとする。例えば、エントリ 2217 に登録された値が「50」の場合、輻輳が発生したパスの負荷分散の割合を、以前の割合の 50% とする。

【0039】

図 5 は、パス管理テーブル 222 のデータ構造例を示す図である。パス管理テーブル 222 は、ホストストレージ装置間通信又はストレージ装置間通信で使用されるパスごとに 1 つのレコードを有する。パス管理テーブル 222 の各レコードは、レコードに対応する個々のパスを識別する為の識別子であるパス ID が登録されるエントリ 2221、当該レコードに対応するパスを使用するリモートコピータスクのタスク ID が登録されるエントリ 2222、当該パスが経由するネットワークのネットワーク ID が登録されるエントリ 2223、当該パス上で輻輳が発生していない時点での割合であるデフォルト割合が登録されるエントリ 2224、当該パスの現時点での割合が登録されるエントリ 2225、当該パス上で発生した輻輳の回数が登録されるエントリ 2226、当該パスのマスタストレージ装置 1 側の IP アドレスであるイニシエータ IP アドレスが登録されるエントリ 2227 および当該パスのリモートストレージ装置 5 側の IP アドレスであるターゲット IP アドレスが登録されるエントリ 2228 を有する。

【0040】

なお第一の実施形態では、リモートストレージ装置 5 が、ウェルノウンポートである「3260」以外の TCP ポート番号を iSCSI 通信の通信ポートに使用する場合、エントリ 2228 には、IP アドレスと前記 TCP ポート番号とを「:」でつなげた文字列がターゲット IP アドレスとして登録されるものとする。ここで、ウェルノウンポートとは、IANA (Internet Assigned Numbers Authority) が、iSCSI 等のアプリケーションレイヤプロトコルに割り当てた TCP ポート番号である。

【0041】

本実施形態では、ストレージ管理装置 2 がストレージ装置間通信で使用されているパスの状況を監視する。ここで、パスを構成するネットワーク装置 3 は、輻輳が発生すると、輻輳の発生をストレージ管理装置 2 へ通知する。輻輳の発生を通知されたストレージ管理装置 2 は、あらかじめ設定された変更率でマスタストレージ装置 1 が使用するパスの割合の変更後の値を計算し、その結果をマスタストレージ装置 1 へ送信する。結果を受信したマスタストレージ装置 1 は、受け取った情報に基づいてパスの割合を変更してストレージ装置間通信を継続する。

【0042】

まず、システム管理者等があらかじめストレージ装置間通信のパス、そのパスの使用率、輻輳時のパスの割合の変更率等の情報をストレージ管理装置 2 及びマスタストレージ装置 1 に設定する際の処理手順について説明する。

【0043】

最初に、第一の実施形態で用いられるグラフィカルユーザインタフェース（以下「GUI」）について説明する。これらの GUI は、CPU 204 が GUI 制御プログラム 213 を実行することによってディスプレイ 205 に表示される。システム管理者等は、文字入力装置 207 およびポインティング装置 206 を用いて、表示された GUI 上で各パラメータを設定する。

【0044】

なお、ディスプレイ 205、文字入力装置 207 およびポインティング装置 206 は、ストレージ管理装置 2 とは別の計算機が有していてもよい。例えば、ストレージ管理装置 2 と I-P-SAN 7 あるいはシリアルケーブルを介して接続されるコンソール用端末がディスプレイ 205 等を有していても良い。この場合、CPU 204 は、GUI 制御プログラム 213 を実行して画面データをコンソール用端末へ送信し、コンソール用端末がディスプレイ 205 に GUI を表示する。さらに、コンソール用端末は、システム管理者等が

文字入力装置 207 とポインティング装置 206 とを用いて設定した各パラメータをストレージ管理装置 2 へ送信する。

【0045】

また、ストレージ管理装置 2 は、第一の実施形態で説明する GUI のかわりに、その GUI と同等の機能を有するコマンドラインインタフェースを備えてもよい。

【0046】

図 6 は、システム管理者等が、ネットワーク装置 3 の情報をストレージ管理装置 2 に登録するために使用するネットワーク登録画面 600 の表示例を示す図である。ネットワーク登録画面 600 は、ネットワーク ID が入力される領域 601、ネットワークを構成するネットワーク装置 3 の情報が入力される領域 602、領域 602 で入力された情報をネットワーク管理テーブル 220 に追加する際に使用されるボタン 605、領域 607 を用いて指定されたネットワーク装置 3 に相当するレコードをネットワーク管理テーブル 220 から削除する際に使用されるボタン 606、既にネットワーク管理テーブル 220 に登録されたネットワーク装置 3 のリストを表示する領域 607 およびネットワーク登録画面 600 を閉じる際に使用されるボタン 618 を有する。

【0047】

さらに、領域 602 は、ネットワーク装置 3 の装置 ID が入力される領域 603 およびネットワーク装置 3 が送信する輻輳発生通知や輻輳回復通知のソース IP アドレスが入力される領域 604 を有する。

【0048】

図 7 は、システム管理者等が、リモートコピータスクの情報およびリモートコピータスクが使用する 1 以上のパスの情報をストレージ管理装置 2 に登録するために使用するリモートコピータスク登録画面 700 の表示例を示す図である。リモートコピータスク登録画面 700 は、タスク ID が入力される領域 701、イニシエータ iSCSI 名が入力される領域 702、イニシエータ LUN が入力される領域 703、ターゲット iSCSI 名が入力される領域 704、ターゲット LUN が入力される領域 705、パスの情報を設定する領域 706、パス間の負荷分散の割合の設定を行う領域 719、これらの領域で指定された情報を登録する際に使用されるボタン 728 および登録を取り消す際に使用されるボタン 729 を有する。

【0049】

さらに、領域 706 は、パスのパス ID が入力される領域 707、当該パスが経由するネットワークのネットワーク ID の一覧を表示してその内の一つを選択する際に使用されるボタン 709、ボタン 709 で選択されたネットワーク ID を表示する領域 708、当該パスのイニシエータ IP アドレスの一覧を表示してその内の一つを選択する際に使用されるボタン 711、ボタン 711 で選択された IP アドレスを表示する領域 710、当該パスのターゲット IP アドレスの一覧を表示してその内の一つを選択する際に使用されるボタン 713、管理者によって入力された又はボタン 713 で選択されたターゲット IP アドレスを表示する領域 712、領域 704 に入力された iSCSI 名に対応する IP アドレスとポート番号とを iSNS (internet Simple Naming Service) サーバや SLP DA (Service Location Protocol Directory Agent) サーバ等のネーム管理サーバから読み出し、ボタン 713 で選択可能にする際に使用されるボタン 714、当該パスの負荷分散の割合が入力される領域 715、領域 707 から領域 715 までで指定された情報をパス管理テーブル 222 に追加する際に使用されるボタン 716、領域 718 を用いて指定されたパスに相当するレコードをパス管理テーブル 222 から削除する際に使用されるボタン 717 および既にパス管理テーブル 222 に登録されたパスのリストを表示する領域 718 を有する。

【0050】

また、領域 719 は、輻輳時にパス間の負荷分散の割合を変更することを指定する際に使用されるボタン 720 およびボタン 720 が選択された時に入力可能となり、輻輳した

パスの割合変更率が百分率で入力される領域 721 を有する。

【0051】

なお、ボタン 714 が指定された時にストレージ管理装置 2 が通信を行うネーム管理サーバの IP アドレスは、ストレージ管理装置 2 にあらかじめ設定されているものとする。

【0052】

以下、GUI 操作で各種情報が登録される時のストレージ管理装置 2 の処理について説明する。なお、以下の処理は、CPU 204 で実行される GUI 制御プログラム 213 によって実行される。

【0053】

図 8 (a) は、システム管理者等が、ネットワーク登録画面 600 のボタン 605 あるいはボタン 606 を指定した時に、ストレージ管理装置 2 が実行するネットワーク登録処理の動作手順を示すフローチャートである。この処理によって、システムの管理者等は、ストレージ管理装置 2 にネットワーク装置 3 の情報を追加 (又は削除) する。

【0054】

まず、ストレージ管理装置 2 は、指定されたのがボタン 605 (追加) かボタン 606 (削除) かを判定する (S801)。ボタン 605 が指定された場合、ストレージ管理装置 2 は、GUI で入力された情報に基づいてネットワーク管理テーブル 220 にレコードを追加する。追加されるレコードのエントリ 2201、エントリ 2202 およびエントリ 2203 には、ボタン 605 が指定された時の領域 601、領域 603 および領域 604 の内容がそれぞれ登録される (S802)。一方、ボタン 606 が指定された場合、ストレージ管理装置 2 は、ボタン 606 が指定された時に、領域 607 で指定されたネットワーク装置 3 の装置 ID がエントリ 2202 の内容と一致する条件で、ネットワーク管理テーブル 220 を検索する (S803)。そして、ストレージ管理装置 2 は、見つかったレコードをネットワーク管理テーブル 220 から削除する (S804)。以上で、ストレージ管理装置 2 はネットワーク登録処理を終了する。

【0055】

図 8 (b) は、システム管理者等が、リモートコピータスク登録画面 700 のボタン 716 あるいはボタン 717 を指定した時に、ストレージ管理装置 2 が実行するパス登録処理の動作手順を示すフローチャートである。システム管理者等は、この処理によって、リモートコピーの 1 つのタスクで使用する複数のパス、パスの割合、輻輳時のパスの割合の変更の可否等の情報をストレージ管理装置 2 に追加する。又、管理者等は、この処理によって、一旦設定したリモートコピーに使用するパスの本数等の変更 (パス数の増加又は減少、パスの割合の変更等) をストレージ管理装置 2 に登録することが出来る。

【0056】

まず、ストレージ管理装置 2 は、指定されたのがボタン 716 (追加) かボタン 717 (削除) かを判定する (S811)。

【0057】

ボタン 716 が指定された場合、ストレージ管理装置 2 は、GUI で入力された情報に基づいてパス管理テーブル 222 にレコードを追加する。ここで追加するレコードのエントリ 2221、エントリ 2222、エントリ 2223 およびエントリ 2224 には、ボタン 716 が指定された時の領域 707、領域 701、領域 708 及び領域 715 の内容がそれぞれ登録される。エントリ 2225 にも、ボタン 716 が指定された時の領域 715 の内容が登録される。エントリ 2226 には「0」が登録される。エントリ 2227 およびエントリ 2228 には、ボタン 716 が指定された時の領域 710 および領域 712 の内容がそれぞれ登録される (S812)。

【0058】

一方、ボタン 717 が指定された場合、ストレージ管理装置 2 は、ボタン 717 が指定された時に、領域 718 で指定されたパスのタスク ID およびパス ID が、それぞれエントリ 2222 およびエントリ 2221 の内容と一致する条件で、パス管理テーブル 222 を検索する (S813)。そして、ストレージ管理装置 2 は、見つかったレコードをパス

管理テーブル 222 から削除する (S814)。以上で、ストレージ管理装置 2 はパス登録処理を終了する。

【0059】

図 8 (c) は、システム管理者等が、リモートコピータスク登録画面 700 を用いて、1 つのリモートコピータスクの情報を登録した時に、ストレージ管理装置 2 が実行するリモートコピータスク登録処理の動作手順を示すフローチャートである。

【0060】

この処理によって、システム管理者等は、1 つのタスクについて、図 8 (b) で示すパスの登録が正しく (パスの割合が 100%) 設定されたかを確認して、そのタスクの情報をストレージ管理装置 2 へ登録することができる。

【0061】

まず、ストレージ管理装置 2 は、領域 701 で入力されたタスク ID がエントリ 2222 の内容と一致する条件で、パス管理テーブル 222 を検索する (S821)。そして、見つかった全てのレコードのエントリ 2225 (割合) を取り出し、割合の和が 100 であることを確認する (S822)。全ての割合の和が 100 ではない場合、ストレージ管理装置 2 は、ディスプレイ 205 にエラーを伝える GUI を表示して、リモートコピータスク登録処理を終了する。この場合、システム管理者等は、再度パスの登録を行う (S825)。

【0062】

全ての割合の和が 100 の場合、ストレージ管理装置 2 は、タスク管理テーブル 221 にレコードを追加する。ここで追加するレコードのエントリ 2211 には領域 701 の内容が、エントリ 2212 には領域 702 の内容が、エントリ 2213 には領域 703 の内容が、エントリ 2214 には領域 704 の内容が、エントリ 2215 には領域 705 の内容が、エントリ 2216 にはボタン 720 がオフの場合には「0」が、ボタン 720 がオンの場合には「1」が、エントリ 2217 にはボタン 720 がオフの場合には「0」が、ボタン 720 がオンの場合には領域 721 の内容が、それぞれ登録される (S823)。

【0063】

最後に、ストレージ管理装置 2 は、S823 で登録されたりモートコピータスクの情報 (タスク ID、イニシエータ iSCSI 名、イニシエータ LUN、ターゲット iSCSI 名およびターゲット LUN) と、このリモートコピータスクで使用する全てのパスの情報 (パス ID、割合、イニシエータ IP アドレスおよびターゲット IP アドレス) とを含むリモートコピー初期化要求を作成し、マスタストレージ装置 1 へ送信する (S824)。以上で、ストレージ管理装置 2 は、リモートコピータスク登録処理を終了する。その後、リモートコピー初期化要求を受信したマスタストレージ装置 1 は、CPU 104 でリモートコピー実行プログラム 111 を実行し、リモートコピー初期化処理を行う。

【0064】

まず、マスタストレージ装置 1 は、受信したリモートコピー初期化要求からリモートコピータスクの情報とパスの情報とをディスク装置 103 に記憶する。そして、マスタストレージ装置 1 は、リモートストレージ装置 5 との間で iSCSI セッションを確立する。この iSCSI セッションの、イニシエータの iSCSI 名は、受信したリモートコピー初期化要求から取り出したイニシエータ iSCSI 名とし、ターゲットの iSCSI 名は、受信したリモートコピー初期化要求から取り出したターゲット iSCSI 名とする。

【0065】

また、マスタストレージ装置 1 は、受信したリモートコピー初期化要求から取り出したパスの情報ごとに、イニシエータ IP アドレスをソース IP アドレスとし、ターゲット IP アドレスをデスティネーション IP アドレスとする TCP コネクションをリモートストレージ装置 5 との間で確立し、すでに確立した iSCSI セッションに割り当てる。

【0066】

最後に、マスタストレージ装置 1 は、確立した iSCSI セッションを用いて初期コピーを実行する。この初期コピーでは、マスタストレージ装置 1 は、受信したリモートコピ

ー初期化要求から取り出したイニシエータLUNで識別されるマスタストレージ装置1の論理ユニットに格納されたデータを、受信したリモートコピー初期化要求から取り出したターゲットLUNで識別されるリモートストレージ装置5の論理ユニットに複製する。

【0067】

尚、本実施形態では、リモートコピー初期化要求に基づいてストレージ管理装置2がマスタストレージ装置1にリモートコピーのタスクやパスの情報の登録とリモートコピーの初期化（初期コピー）及び開始の指示をまとめて行う形態とした。しかし、ストレージ管理装置2は、システム管理者等の情報登録の際に、リモートコピーのタスクやパスの情報のマスタストレージ装置2への送信のみを行い、その後、システム管理者（あるいはホスト4の使用者）からのリモートコピーの開始（又は準備の開始）指示を受け取った際に、ストレージ管理装置2（又はホスト4）が、マスタストレージ装置1へリモートコピーの初期化処理を指示しても良い。この場合、マスタストレージ装置1は、リモートコピーのパスの情報を受け取った際にはその情報をディスク装置103に格納する処理だけを行う。その後、マスタストレージ装置1は、リモートコピーの初期化処理を指示された際に、格納された情報に基づいてセッションの確立、初期化コピー等の処理を行う。

【0068】

リモートコピー初期化要求に基づく処理の終了後（又はリモートコピー初期化要求の受領後）、マスタストレージ装置1はリモートコピーを開始する。リモートコピー実行中に、ストレージ管理装置2は、リモートコピーで使用されるパスの状態を監視している。

【0069】

次に、第一の実施形態においてマスタストレージ装置1がリモートコピーを実行している際にネットワーク装置3で輻輳が発生した場合の各装置間の通信シーケンスについて説明する。

【0070】

図9は、マスタストレージ装置1とリモートストレージ装置5との間でリモートコピーを実行中にネットワーク装置3にて輻輳が発生し、その後輻輳が回復するまでの、装置間での通信シーケンス例を示す図である。

【0071】

まず、マスタストレージ装置1が、ホスト4からデータ書き込み要求を受信すると、書き込まれたデータをリモートストレージ装置5へ転送する。ここで、マスタストレージ装置1は、リモートコピー初期化処理においてストレージ管理装置2から受信したパスの割合に従って、データ転送に使用するTCPコネクションを選択する。例えば、マスタストレージ装置1が、リモートストレージ装置5との通信に、二つのパスを使用し、かつそれぞれの割合がどちらも50%であった場合、マスタストレージ装置1は、S901のデータ転送を第一のパスに対応するTCPコネクションを用いて実行し、S902のデータ転送を第二のパスに対応するTCPコネクションを用いて実行する（S901、S902）。

【0072】

その後ネットワーク装置3にて輻輳が発生すると（S903）、そのネットワーク装置3は輻輳発生通知をストレージ管理装置2へ送信する（S904）。

【0073】

輻輳発生通知を受信したストレージ管理装置2は、パス管理テーブル更新処理および割合計算処理を実行し、変更後のパス間の割合を得る。そして、ストレージ管理装置2は、マスタストレージ装置1へパスの割合変更要求を送信する（S905）。割合変更要求を受信したマスタストレージ装置1は、割合変更要求から変更後のパス間の割合を取り出してディスク装置103に記憶することで変更前のパス間の割合の値を変更後の値に変更した後、割合変更応答をストレージ管理装置2へ送信する（S906）。

【0074】

その後、マスタストレージ装置1は、変更後のパス間の割合を用いて、リモートストレージ装置5へコピーするデータを転送する。例えば、パス間の割合が75%と25%とに

変更されたら、マスタストレージ装置 1 は、データ転送の際、第一のパスに対応する TCP コネクションを 4 回に 3 回の割合で使用し、第二のパスに対応する TCP コネクションを 4 回に 1 回の割合で使用する (S907)。

【0075】

その後、ネットワーク装置 3 にて発生した輻輳が回復すると (S908)、ネットワーク装置 3 は輻輳回復通知をストレージ管理装置 2 へ送信する (S909)。輻輳回復通知を受信したストレージ管理装置 2 は、パス管理テーブル更新処理および割合計算処理を実行し、変更後のパス間の割合を得る。

【0076】

そして、ストレージ管理装置 2 は、マスタストレージ装置 1 へパスの割合変更要求を送信する (S910)。割合変更要求を受信したマスタストレージ装置 1 は、パス間の割合を変更した後、割合変更応答をストレージ管理装置 2 へ送信する (S911)。そして、マスタストレージ装置 1 は、変更後のパス間の割合を用いて、リモートストレージ装置 5 へデータを転送する。

【0077】

第一の実施形態では、ネットワーク装置 3 は、SNMP (Simple Network Management Protocol) Trap を用いて、輻輳発生通知や輻輳回復通知を、WAN8 および IP-SAN7 を介してストレージ管理装置 2 へ送信することを想定している。ここで、SNMP Trap を用いた輻輳発生通知および輻輳回復通知には、それぞれ輻輳発生および輻輳回復を表す OID (Object Identifier) が含まれる。しかしネットワーク装置 3 はこれ以外のプロトコルを用いて輻輳の発生等をストレージ管理装置 2 に通知してもよい。

【0078】

次に、図 9 で説明したパス管理テーブル更新処理や割合計算処理の詳細な処理手順について説明する。ここでの処理手順の概略は以下のとおりとなる。まず、ストレージ管理装置 2 が輻輳の発生したネットワーク装置 3 に関連するパスを特定する。次に、ストレージ管理装置 2 は特定されたパスを使用するタスクを選択する。その後、ストレージ管理装置 2 は、選択されたタスクでパス間の割合変更が必要な場合に変更後の値を計算し、その結果をタスクを実行しているストレージ装置へ割合変更要求として送信する。

【0079】

図 10 および図 11 は、ストレージ管理装置 2 におけるパス管理テーブル更新処理の動作手順を示すフローチャートである。なお、以下の処理は、CPU204 で割合計算プログラム 211 が実行されることによって実行される。

【0080】

パス管理テーブル更新処理を開始したストレージ管理装置 2 は、まず、パス管理テーブル 220 のエントリ 2203 に登録された値が受信した輻輳発生通知を送信したネットワーク装置 3 の IP アドレスと一致するという条件でネットワーク管理テーブル 220 を検索する。そして、ストレージ管理装置 2 は、該当するレコードのネットワーク ID (エントリ 2201) を読み出す (S1001)。

【0081】

次にストレージ管理装置 2 は、エントリ 2223 に登録された値が S1001 で読み出したネットワーク ID と一致するという条件でパス管理テーブル 222 を検索する。そして、ストレージ管理装置 2 は、検索条件に合致するレコードを全て読み出す (S1002)。

【0082】

次にストレージ管理装置 2 は、受信した通知から取り出した OID から、通知された内容が輻輳発生か輻輳回復かを判断する (S1003)。輻輳発生の場合、ストレージ管理装置 2 は S1002 で読み出した全てのレコードの輻輳回数 (エントリ 2226) の値に 1 を加算する (S1004)。一方、輻輳回復の場合、ストレージ管理装置 2 は S1002 で読み出した全てのレコードの輻輳回数 (エントリ 2226) の値から 1 を減算する (

S1005)。

【0083】

輻輳回数の加減算をした後、ストレージ管理装置2はS1002で読み出した各レコードのタスクID (エントリ2222) を読み出し、タスクIDのリストを作成する。この際、ストレージ管理装置2は重複しているタスクIDをリストから除外する (S1006)。

【0084】

(ここから図11) 次にストレージ管理装置2は、S1006で作成したタスクIDリストの先頭のタスクIDを選択する (S1101)。そしてストレージ管理装置2は、エントリ2211に登録された値がS1101で選択したタスクIDと一致するという条件で、タスク管理テーブル221を検索する。その後ストレージ管理装置2は、検索条件に合致するレコードの割合変更フラグ (エントリ2216) と割合変更率 (エントリ2217) とを読み出す (S1102)。

【0085】

次にストレージ管理装置2は、読み出した割合変更フラグの値が「0」か「1」かを判定する (S1103)。割合変更フラグの値が「1」の場合、ストレージ管理装置2は、エントリ2222に登録された値がS1101で読み出したタスクIDと一致するという条件でパス管理テーブル222を検索する。そしてストレージ管理装置2は、検索条件に合致するレコードを全て読み出し (S1104)、図12で説明する割合計算処理を実行する (S1105)。この割合計算処理の結果に基づき、ストレージ管理装置2は、S1104で読み出した各レコードの割合 (エントリ2225の値) を更新する (S1106)。S1106の処理の終了後又はS1103で割合変更フラグが「0」と判断された場合、ストレージ管理装置2は、S1102からS1106までの処理をタスクIDリストに含まれる全てのタスクIDについて実行し (S1107、S1108)、パス管理テーブル更新処理を終了する。

【0086】

図12は、ストレージ管理装置2における割合計算処理の動作手順を示すフローチャートである。なお、以下の処理は、CPU204で割合計算プログラム211が実行されることによって実行される。割合計算処理では、輻輳が発生したパスに割り当てられているパス間の割合を指定された変更率で減少させ、その減少分を他のパスに振り分ける処理を行う。また、全てのパスの輻輳回数が一致した場合 (タスクで使用する全てのパスで同じ回数の輻輳が起こった場合に該当する) および全てのパスで輻輳が回復した場合には、割合計算処理ではパス間の割合をデフォルト値に戻す処理が行われる。

【0087】

まずストレージ管理装置2は、図11のS1104で読み出した全てのレコードの輻輳回数 (エントリ2226) の値が一致するか否かを調べる (S1201)。全ての輻輳回数が一致する場合、ストレージ管理装置2は図11のS1104で読み出した全てのレコードのパス間の割合 (エントリ2225) をデフォルト割合 (エントリ2224) の内容で上書きし (S1202)、割合計算処理を終了する。

【0088】

一方S1104で読み出した全てのレコードの中に一つでも異なる輻輳回数の値がある場合、ストレージ管理装置2は全てのレコードの中の輻輳回数の最小値を調べる。この結果得た値を最小輻輳回数という (S1203)。次にストレージ管理装置2は、図11のS1104で読み出した各レコードに対応するパスの相対輻輳回数を、以下の式に基づいて計算する (S1204)。

相対輻輳回数 = 輻輳回数 - 最小輻輳回数

その後ストレージ管理装置2は、相対輻輳回数が0より大きいパスについて、パス間の割合と差分を以下の式に基づいて計算する (S1205)。ただし、「^」はべき乗演算を表す。

パス間の割合 = デフォルト割合 × (割合変更率 ^ 相対輻輳回数)

差分＝パス間の割合－デフォルト割合

またストレージ管理装置 2 は、相対輻輳回数が 0 であるパスについて、パス間の割合を以下の式に基づいて計算する (S 1 2 0 6)。

割合の和＝相対輻輳回数が 0 のパスのデフォルト割合の和

パス間の割合＝デフォルト割合＋S 1 2 0 5 で計算した全ての差分の和×デフォルト割合／割合の和

上述の割合等を計算したストレージ管理装置 2 は、割合計算処理を終了する。

【0089】

上記の割合計算処理について、例えば、ストレージ装置間で行われる 1 つのリモートコピータスクで 3 つのパスが使用されており、かつ第一、第二および第三のパスの割合がそれぞれ 40 (%)、30 (%) および 30 (%) であり、輻輳時の割合変更率が 50 (%) の場合を考える。この場合、第一のパス上で輻輳が発生したら、上記の割合計算処理を行うと、第一のパスの割合は 20 (%)、残りのパスの割合はそれぞれ 40 (%) となる。

【0090】

以上で第一の実施形態を説明した。第一の実施形態によると、マスタストレージ装置 1 がリモートストレージ装置 5 との間の複数のパスを用いて i S C S I を用いたリモートコピーを実行する際に、あるパス上のネットワーク装置 3 で輻輳が発生した時、ネットワーク装置 3 がストレージ管理装置 2 へ輻輳の発生を通知する。通知されたストレージ管理装置 2 は、それを契機に輻輳発生の結果として送信レートが低下したパスの使用率（ここでは割合と称している）を下げるよう、マスタストレージ装置 1 に指示する。これにより、例えばリモートコピー等におけるストレージ装置間のデータ転送のスループット低下を最小限に抑えることができる。

【0091】

なお、第一の実施形態では、マスタサイト 11 にあるマスタストレージ装置 1 と、リモートサイト 12 にあるリモートストレージ装置 5 との間のリモートコピーに本発明を適用したが、1 つのサイト（例えば、マスタサイト 11）にある二つのストレージ装置間のストレージ装置間通信（例えばリモートコピー）にも本発明は適用可能である。この場合、ネットワーク装置 3 は I P - S A N 7 を構成する機器となる。

【実施例 2】

【0092】

次に、第二の実施形態について第一の実施形態と相違する部分に限って説明する。第一の実施形態では、ネットワーク装置 3 がストレージ管理装置 2 に S N M P T r a p 等のプロトコルを用いて輻輳発生通知と輻輳回復通知とを送信していた。しかし、この通知を行うためには S N M P の標準仕様に手を加える必要があり、必ずしも全てのネットワーク装置 3 が実施可能とは限らない。

【0093】

一方、ネットワーク装置 3 は一般に、廃棄したパケット数を記憶し外部からの要求に応じてその廃棄パケット数の情報を外部の装置に送信することができる。そこで本実施形態では、ストレージ管理装置 2 が定期的にネットワーク装置 3 から廃棄パケット数の情報を読み出し、前回読み出した値と比較して廃棄パケット数が増えていたら輻輳が発生したとみなす。一方、輻輳発生とみなしてから一定期間経過した後、ストレージ管理装置 2 が廃棄パケット数を読み出し、前回読み出した値と比較して廃棄パケット数が増えていなかったら輻輳が回復したとみなす。これにより、輻輳発生通知および輻輳回復通知を送信できないネットワーク装置 3 を有するシステムで、輻輳の発生によるパス間の割合の変更を行うようにする。

【0094】

図 13 (a) は、第二の実施形態におけるストレージ管理装置 2 の構成例を示した図である。ストレージ管理装置 2 のディスク装置 203 には、第一の実施形態のネットワーク管理テーブル 220 の代わりにネットワーク管理テーブル 223 が格納される。ネットワ

ーク管理テーブル 223 の各レコードは、第一の実施形態で説明した各エントリに加え、ネットワーク装置 3 がバッファ領域不足のために廃棄したパケット数を表す輻輳廃棄パケット数が登録されるエントリ 2204 を有する。

【0095】

ストレージ管理装置 2 の主メモリ 201 には、第一の実施形態で説明した輻輳受信プログラム 210 の代わりに、ネットワーク管理テーブル 223 に登録された全てのネットワーク装置 3 から定期的に廃棄パケット数の情報を含む統計データを読み出す際に CPU 204 で実行される統計データ読出プログラム 214 が格納される。

【0096】

図 13 (b) は、第二の実施形態におけるネットワーク装置 3 の構成例を示した図である。ネットワーク装置 3 の主メモリ 301 には、第一の実施形態で説明した輻輳状態通知プログラム 311 の代わりに、各種統計データを記憶する統計データ記憶領域 313 および各種統計データを統計データ記憶領域 313 に書き込む際に CPU 304 で実行される統計データ記憶プログラム 312 が格納される。

【0097】

統計データ記憶領域 313 に記憶される統計データには、ネットワーク装置 3 が受信したパケットのうち廃棄したパケットの総数（以下「廃棄パケット総数」という）、廃棄パケット総数のうちパケットのフォーマットエラーのために廃棄したパケット数（以下「エラー廃棄パケット数」という）および廃棄パケット総数のうちネットワーク装置 3 が対応できないプロトコルのパケットであったため廃棄したパケット数（以下「プロトコル不正廃棄パケット数」という）が挙げられる。

【0098】

輻輳廃棄パケット数は以下の式で表される。

輻輳廃棄パケット数 = 廃棄パケット総数 - エラー廃棄パケット数 - プロトコル不正廃棄パケット数

本実施形態におけるネットワーク登録処理は、ネットワーク管理テーブル 220 がネットワーク管理テーブル 223 に置き換わる点および図 8 (a) の S802 にて、エントリ 2204 にレコード追加時点での輻輳廃棄パケット数が登録される点を除き、第一の実施形態と同様である。ここで、ストレージ管理装置 2 は、追加するレコードに対応するネットワーク装置 3 へ統計データ要求を送信し、廃棄パケット総数、エラー廃棄パケット数およびプロトコル不正廃棄パケット数といった統計データを要求する。そして、ストレージ管理装置 2 は、ネットワーク装置 3 から要求された統計データを含む統計データ応答を受信し、その統計データ応答の内容に基づき輻輳廃棄パケット数を計算する。これにより、ストレージ管理装置 2 は、レコード追加時点での輻輳廃棄パケット数を得る。

【0099】

図 14 は、第二の実施形態におけるシステム内の各装置間の通信シーケンス例を示す図である。第一の実施形態と同様、マスタストレージ装置 1 はストレージ管理装置 2 からの情報に基づいてリモートコピーを行っているとする。まず、マスタストレージ装置 1 が、デフォルトのパス間の割合に従って、リモートストレージ装置 5 へコピーするデータの転送を行う (S1401、S1402)。ストレージ管理装置 2 は、定期的にネットワーク装置 3 に対して統計データ要求を送信し、廃棄パケット総数、エラー廃棄パケット数およびプロトコル不正廃棄パケット数といった統計データを要求する (S1403)。

【0100】

統計データ要求を受信したネットワーク装置 3 は、要求された統計データを含む統計データ応答を送信する (S1404)。統計データ応答を受信したストレージ管理装置 2 は、統計データ応答の内容に基づき輻輳廃棄パケット数を計算し、ネットワーク管理テーブル 223 の当該ネットワーク装置 3 に対応するレコードのエントリ 2204 に登録された値と比較する。これらの値が一致した場合、ストレージ管理装置 2 は、そのネットワーク装置 3 で輻輳は発生していないとみなす。本シーケンスでは、この時点ではこれらの値が一致したとする。

【0101】

その後ネットワーク装置3で輻輳が発生するとする(S1405)。ストレージ管理装置2は輻輳発生後の所定のタイミング(又は一定間隔)で統計データ要求を送信し(S1406)、統計データ応答を受信する(S1407)。その後、ストレージ管理装置2は、統計データ応答の内容に基づき輻輳廃棄パケット数を計算し、ネットワーク管理テーブル223の当該ネットワーク装置3に対応するレコードのエントリ2204に登録された値と比較する。この時、輻輳が既に発生しているので、輻輳廃棄パケット数がエントリ2204に登録された値を上回る。

【0102】

この場合、ストレージ管理装置2は、そのネットワーク装置3に輻輳が発生したとみなし、ネットワーク管理テーブル223の当該レコードのエントリ2204に輻輳廃棄パケット数を書き込む。更にストレージ管理装置2は、図10、図11および図12で説明したパス管理テーブル更新処理および割合計算処理を行う。その結果、パスの変更後の割合を得たストレージ管理装置2は、マスタストレージ装置1へパスの割合変更要求を送信し(S1408)、割合変更応答を受信する(S1409)。その後、マスタストレージ装置1は変更後のパス間の割合を用いて、リモートストレージ装置5へコピーするデータを送信する(S1410)。

【0103】

その後、ネットワーク装置3にて発生していた輻輳が回復するとする(S1411)。所定のタイミング(例えば定期的に)でストレージ管理装置2が統計データ要求を送信し(S1412)、統計データ応答を受信する(S1413)。統計データ応答を受信したストレージ管理装置2は、統計データ応答の内容に基づき輻輳廃棄パケット数を計算し、ネットワーク管理テーブル223の当該ネットワーク装置に対応するレコードのエントリ2204に登録された値と比較する。この場合、既に輻輳が回復しているので、これらの値が一致する。この場合輻輳が回復したとみなし、ストレージ管理装置2は、図10、図11および図12で説明したパス管理テーブル更新処理、割合計算処理を行う。この処理の結果パスの変更後のパス間の割合を得たストレージ管理装置2は、マスタストレージ装置1へパスの割合変更要求を送信し(S1414)、割合変更応答を受信する(S1415)。

【0104】

なお、上記の通信シーケンス例では、ストレージ管理装置2がS1407で統計データ応答を受信し輻輳が発生したとみなした後、S1413で統計データ応答を受信し輻輳廃棄パケット数とエントリ2204に登録された値とを比較して一致した時に輻輳が回復したとみなしたが、輻輳廃棄パケット数とエントリ2204に登録された値とが連続して何回一致すれば、輻輳が回復したとみなすか、についてはシステム管理者等が任意に設定できるようにしてもよい。この回数は、例えば、WAN8で輻輳が発生してから回復するまでの平均時間を、ストレージ管理装置2が統計データ要求を送信する間隔で割った結果を切り上げた値とすればよい。

【0105】

以上で第二の実施形態を説明した。第二の実施形態によると、ストレージ管理装置2が定期的にネットワーク装置3から廃棄パケットに関する統計データを読み出して輻輳発生や輻輳回復を判断することにより、割合の変更の指示等を第一の実施形態と同様に行うことができる。

【実施例3】**【0106】**

次に、第三の実施形態を第一の実施形態と相違する部分に限って説明する。第三の実施形態では、ネットワーク装置3は、ECN(Explicit Congestion Notification)というRFC3168で規定された標準仕様を用いて、マスタストレージ装置1に輻輳発生と輻輳回復とを直接通知する。従って、第一および第二の実施形態と比較して、ストレージ管理装置2と各装置間での通信量が削減されるという利

点がある。なおネットワーク装置 3 とマスタストレージ装置 1 との間で使用される通信の仕様は E C N に限る必要は無い。

【0107】

図 15 (a) は、第三の実施形態におけるマスタストレージ装置 1 の構成例を示した図である。マスタストレージ装置 1 の主メモリ 101 には、第一の実施形態で説明した割合変更要求受信プログラム 111 の代わりに、ストレージ管理装置 2 からシステム管理者等が入力した設定内容を受信した際に C P U 104 で実行される設定受信プログラム 113、ネットワーク装置 3 から E C N による輻輳通知や輻輳回復を受信した際に C P U 104 で実行される輻輳受信プログラム 114 およびパス間の負荷分散の割合を計算する際に C P U 104 で実行される割合計算プログラム 115 が格納される。

【0108】

また、ディスク装置 103 には、タスク管理テーブル 221 およびパス管理テーブル 222 が格納される。つまり、第一および第二の実施形態においてストレージ管理装置 2 で行われていた割合計算等の処理がマスタストレージ装置 1 にて行われることになる。

【0109】

図 15 (b) は、第三の実施形態におけるストレージ管理装置 2 の構成例を示した図である。本実施形態においては、ストレージ管理装置 2 は単なる設定情報の入力端末に過ぎない。従って、ストレージ管理装置 2 の主メモリ 201 には、G U I 制御プログラム 213 およびシステム管理者等が入力した設定内容をマスタストレージ装置 1 へ送信する際に C P U 204 で実行される設定送信プログラム 215 が格納される。ディスク装置 203 には、本発明に関わるいずれのテーブルも格納されない。

【0110】

第三の実施形態では、システム管理者等がディスプレイ 205 に表示されたりリモートコピータスク登録画面 700 上で、文字入力装置 207 およびポインティング装置 206 を用いて各パラメータを設定する。その後、システム管理者等がボタン 716、ボタン 717 あるいはボタン 728 を指定すると、設定送信プログラム 215 が C P U 204 で実行され、設定内容がマスタストレージ装置 1 に送信される。なお、第三の実施形態では、リモートコピータスク登録画面 700 の領域 708 およびボタン 709 は不要である。

【0111】

一方、設定内容を受信したマスタストレージ装置 1 は、設定受信プログラム 113 を C P U 104 で実行し、ストレージ管理装置 2 から受信した設定内容に基づき、タスク管理テーブル 221 およびパス管理テーブル 222 にレコードを登録あるいはテーブルからレコードを削除する。登録されるレコードの内容は、パス管理テーブル 222 のエントリ 2223 には何も登録されない点を除き、図 8 (b) および (c) で説明したそれぞれパス登録処理およびリモートコピータスク登録処理と同様である。なお、第三の実施形態ではネットワーク装置 3 をストレージ管理装置 2 で管理する必要が無いため、ネットワーク登録画面 600 は不要である。

【0112】

図 16 は、第三の実施形態における各装置間の通信シーケンス例を示す図である。ここでは、マスタストレージ装置 1 からリモートストレージ装置 5 へリモートコピーが実施されているとする。尚、リモートコピーの初期化要求や開始指示等は第一の実施形態と同じである。又、第三の実施形態で使用する E C N は T C P の拡張仕様であるため、図 16 には、T C P レイヤの通信まで記す。

【0113】

まず、マスタストレージ装置 1 が、デフォルトのパス間の割合に従って、リモートストレージ装置 5 へコピーするデータを含むパケットを送信する (S1601)。なお、マスタストレージ装置 1 は、E C N を用いることを示すために、S1601 で送信するパケットの I P ヘッダの E C T (E C N C a p a b l e T r a n s p o r t) フラグを O N にする。以降、第三の実施形態において、マスタストレージ装置 1 およびリモートストレージ装置 5 は、送信する全てのパケットの I P ヘッダの E C T フラグを、同様に O N にす

る。

【0114】

ネットワーク装置3は、パケットを受信すると、それをそのままリモートストレージ装置5へ転送する。パケットを受信したリモートストレージ装置5は、受信確認応答（以下「ACK」という）をマスタストレージ装置1へ送信する（S1602）。ネットワーク装置3は、ACKを受信すると、そのままACKをマスタストレージ装置1へ転送する。

【0115】

その後、ネットワーク装置3にて輻輳が発生した場合（S1603）、マスタストレージ装置1がリモートストレージ装置5へコピーするデータを含むパケットを送信すると（S1604）、そのパケットを受信したネットワーク装置3は、受信したパケットのIPヘッダのECTフラグがONであることを確認し、そのパケットのCE（Congestion Experienced）フラグをONにした上で、そのパケットをリモートストレージ装置5へ転送する。ここで、CEフラグとは、ネットワーク装置3が輻輳の有無をエンドポイントに伝えるためのIPヘッダのフラグであり、RFC3168で規定されている（S1605）。

【0116】

CEフラグがONになったパケットを受信したリモートストレージ装置5は、ECE（ECN Echo）フラグをONにしたACKをマスタストレージ装置1へ送信する。ここで、ECEフラグとは、TCP/IP通信を行うエンドポイント間で輻輳の有無に関する情報を交換するためのTCPヘッダのフラグであり、RFC3168で規定されている（S1606）。ネットワーク装置3は、そのACKをマスタストレージ装置1へ転送する。

【0117】

TCPヘッダのECEフラグがONになったACKを受信したマスタストレージ装置1は、WAN8内で輻輳が発生したとみなし、図17で説明するパス管理テーブル更新処理および図12で説明した割合計算処理を実行する。その結果パスの変更後の割合を得たマスタストレージ装置1は、パス間の割合を変更した上で、リモートストレージ装置5へ、コピーするデータを含むパケットを送信する（S1607）。

【0118】

この際、マスタストレージ装置1は、S1607で送信するパケットのTCPヘッダのCWR（Congestion Window Reduced）フラグをONにする。ここで、CWRフラグとは、TCP/IP通信を行うエンドポイント間で輻輳への対策を実行したことを伝えるためのTCPヘッダのフラグであり、RFC3168で規定されている。

【0119】

このパケットを受信したネットワーク装置3は、まだ輻輳が継続しているため、CEフラグをONにしてパケットをリモートストレージ装置5へ転送する（S1608）。リモートストレージ装置5が、CWRフラグ及びCEフラグとがONであるパケットを受信すると、S1606と同様に、ECEフラグをONにしたACKをマスタストレージ装置1に送信する（S1609）。

【0120】

その後、ネットワーク装置3にて発生していた輻輳が回復し（S1610）、マスタストレージ装置1がリモートストレージ装置5へコピーするデータを含むパケットを送信すると（S1611）、そのパケットを受信したネットワーク装置3は、CEフラグをOFFのまま、そのパケットをリモートストレージ装置5へ転送する（S1612）。

【0121】

リモートストレージ装置5は、CEフラグがOFFのパケットを受信すると、ECEフラグがOFFのACKをマスタストレージ装置1へ送信する（S1613）。ECEフラグがOFFのACKを受信したマスタストレージ装置1は、WAN8の輻輳が回復したとみなし、図17で説明するパス管理テーブル更新処理および図12で説明した割合計算処

理を実行する。その結果、マスタストレージ装置 1 は、パス間の割合を元に戻してデータ転送を行う。

【0122】

図 17 は、第三の実施形態におけるマスタストレージ装置 1 のパス管理テーブル更新処理の動作手順を示すフローチャートである。なお、以下の処理は、CPU 104 で割合計算プログラム 115 が実行されることによって実行される。

【0123】

マスタストレージ装置 1 は、まずエントリ 2227 に登録された値が受信したパケットの宛先（デスティネーション）IP アドレスに一致し、かつエントリ 2228 に登録された値が受信したパケットのソース IP アドレスに一致するという条件でパス管理テーブル 222 を検索し、検索条件に該当するレコードを読み出す。ここで、受信したパケットの宛先 IP アドレスはマスタストレージ装置 1 の IP アドレスであり、ソース IP アドレスはリモートストレージ装置 5 の IP アドレスである（S1701）。

【0124】

次に、マスタストレージ装置 1 は、受信したパケットの内容が輻輳発生を示すか輻輳回復を示すかを判定する。具体的には、受信したパケットの TCP ヘッダの ECN フラグが ON であれば輻輳発生を示し、あるいは OFF であれば輻輳回復であると判定する（S1702）。輻輳発生の場合、マスタストレージ装置 1 は、S1701 で読み出したレコードの輻輳回数（エントリ 2226）の値に 1 加算する（S1703）。一方輻輳回復の場合、マスタストレージ装置 1 は、S1701 で読み出したレコードの輻輳回数（エントリ 2226）の値から 1 減算する（S1704）。

【0125】

輻輳回数の加減算の後、マスタストレージ装置 1 は、当該レコードからタスク ID（エントリ 2222）を読み出す（S1705）。その後、マスタストレージ装置 1 は、エントリ 2211 に登録された値が S1705 で読み出したタスク ID と一致するという条件でタスク管理テーブル 221 を検索し、検索条件に合致するレコードの割合変更フラグ（エントリ 2216）と割合変更率（エントリ 2217）とを読み出す（S1706）。

【0126】

次に、マスタストレージ装置 1 は、読み出した割合変更フラグの値が「0」か「1」かを判定する（S1707）。割合変更フラグが「0」の場合、マスタストレージ装置 1 は、パス管理テーブル更新処理を終了する。一方、割合変更フラグが「1」の場合、マスタストレージ装置 1 は、エントリ 2222 に登録された値が S1705 で読み出したタスク ID と一致するという条件でパス管理テーブル 222 を検索し、検索条件に合致したレコードを全て読み出す（S1708）。

【0127】

その後、マスタストレージ装置 1 は、図 12 で説明した割合計算処理を実行する（S1709）。この割合計算処理の結果に基づき、マスタストレージ装置 1 は S1708 で読み出した各レコードの割合（エントリ 2225）の値を更新し（S1710）、パス管理テーブル更新処理を終了する。

【0128】

以上で第三の実施形態を説明した。第三の実施形態によると、ネットワーク装置 3 がマスタストレージ装置 1 へ輻輳発生と輻輳回復とを直接通知することにより、第一の実施形態と比較してストレージ管理装置 2 の CPU 負荷やストレージ管理装置 2 と他の装置との間のトラフィック量の増大を防止できる。

【実施例 4】

【0129】

次に、第四の実施形態を第三の実施形態と相違する部分に限って説明する。第三の実施形態では、ネットワーク装置 3 が、ECN を用いてマスタストレージ装置 1 へ輻輳発生と輻輳回復とを通知する。本実施形態では、マスタストレージ装置 1 自体が輻輳発生や輻輳回復を検出する構成とする。

【0130】

第四の実施形態では、マスタストレージ装置1が、リモートコピーに使用するTCPコネクションにおいてリモートストレージ装置5から送信されるACKがタイムアウト、すなわち一定期間経過しても返ってこない場合や、同じシーケンス番号を持つACKを3つ以上受信した場合に輻輳発生とみなす。これらの動作は、それぞれRFC793およびRFC2581が規定するTCPの仕様に基づいている。

【0131】

マスタストレージ装置1は、上記の動作により輻輳発生を検出した場合、図10、図11および図12で説明したパス管理テーブル更新処理および割合計算処理を行い、その結果得た変更後のパス間の割合を用いて、リモートコピーを実行する。

【0132】

一方、どのような契機を用いて輻輳回復とみなすかは、RFCで規定されていない。そこで、第四の実施形態では、例えば、輻輳発生からあらかじめシステム管理者等が設定した時間が経過した時に輻輳回復とみなすこととする。あるいは、TCPコネクションの輻輳ウィンドウサイズがあらかじめシステム管理者等が設定したサイズを超過した時に輻輳回復とみなしてもよい。マスタストレージ装置1は、上記の動作により輻輳回復を検出した場合、図10、図11および図12で説明したパス管理テーブル更新処理および割合計算処理を行い、その結果得た変更後のパス間の割合を用いて、リモートコピーを実行する。

【0133】

以上で、第四の実施形態を説明した。第四の実施形態によると、マスタストレージ装置1が、自律的に輻輳発生および輻輳回復を検出することにより、ネットワーク装置3に依存せずにパス間の割合を変更することができる。

【実施例5】**【0134】**

次に、第五の実施形態を第一の実施形態と相違する部分に限って説明する。第五の実施形態は、ホストストレージ装置間通信、すなわちホストとストレージ装置との間で実行されるデータの読み出しや書き込み（以下「ホストI/O」ともいう）に本発明を適用したシステムに関する。第五の実施形態では、ホストが、ストレージ装置と1つ以上のTCPコネクションから成る1つのiSCSIセッションを確立してホストI/Oを実行するものとする。TCPコネクションを用いて送信されるパケットが通る経路がパスに相当する。

【0135】

図18は、第五の実施形態のシステム構成例を示した図である。システムは、データを記憶するストレージ装置15、ストレージ装置15からデータを読み出したり、データを書き込むホスト16、端末6、ストレージ管理装置2、ホスト16とストレージ装置15とを接続するネットワーク装置3を有するIP-SAN7、端末6とホスト16とを接続するLAN9、ホスト16、ストレージ装置15並びにストレージ管理装置2およびネットワーク装置3とを接続する管理用ネットワーク13を有する。

【0136】

また、ホスト16およびストレージ装置15とIP-SAN7とは、通信線10で接続されている。同様に、端末6およびホスト16とLAN9とは、通信線10で接続されている。さらに、ホスト16、ストレージ装置15、ストレージ管理装置2およびネットワーク装置3と管理用ネットワーク13とは、通信線14で接続されている。なお、第一の実施形態でも述べた通り、IP-SAN7はLAN9を兼ねても良い。

【0137】

図19はホスト16の構成例を示した図である。ホスト16は、主メモリ601、通信線602、ディスク装置603、CPU604、ディスプレイ605、ポインティング装置606、文字入力装置607および複数のNIF608を有する。1つの以上のNIF608は通信線14経由で管理用ネットワーク13に、残りの1つ以上のNIF608は

通信線 10 経由で I P - S A N 7 に接続される。

【0138】

主メモリ 601 には、ストレージ管理装置 2 から割合変更要求を受信する際に C P U 604 で実行される割合変更要求受信プログラム 610 およびストレージ装置 15 からデータを読み出したり、データを書き込む時に使用するパスを決定する際に C P U 604 で実行されるマルチパス制御プログラム 611 が格納される。

【0139】

ストレージ装置 15 は、単体の記憶装置又は複数の記憶装置を有する記憶装置システムである。

ストレージ管理装置 2 およびネットワーク装置 3 の構成は第一の実施形態と同様である。また、端末 6 は一般的な計算機であり、C P U、主メモリ、入出力装置および他の装置と通信線 10 を介して接続するための N I F 等を有する。

【0140】

第五の実施形態における各種テーブルのデータ構造、G U I および G U I 操作で各種情報が登録される際のストレージ管理装置 2 の処理は第一の実施形態と同様である。

【0141】

第五の実施形態における各装置間の通信シーケンスは、マスタストレージ装置 1 がホスト 16 に置き換わる点、リモートストレージ装置 5 がストレージ装置 15 に置き換わる点とを除き、第一の実施形態と同様である。

【0142】

つまり、マスタストレージ装置 1 に代わってホスト 16 がストレージ管理装置 2 からパスの割合変更要求を受けて割合の変更を行うこととなる。

【0143】

第五の実施形態におけるパス管理テーブル更新処理および割合計算処理の動作手順は、第一の実施形態と同様である。

【0144】

以上で、第五の実施形態を説明した。第五の実施形態によると、ホスト 16 が、ストレージ装置 15 との間の複数のパスを用いて i S C S I プロトコルに基づいたホスト I / O を実行するシステムにおいて、あるパス上のネットワーク装置 3 で輻輳が発生した時、ネットワーク装置 3、ストレージ管理装置 2 およびホスト 16 とが輻輳の情報をやり取りすることで、輻輳発生の結果送信レートが低下したパスの負荷分散の割合を下げるができる。これにより、第一の実施形態と同様に、通信に係るスループット低下を最小限に抑えることができる。

【実施例 6】

【0145】

次に、第六の実施形態について第五の実施形態と相違する部分に限って説明する。第六の実施形態では、第二の実施形態と同様に、ストレージ管理装置 2 が定期的にネットワーク装置 3 から廃棄パケット数を読み出すことで、輻輳発生および輻輳回復を検出する。

【0146】

第六の実施形態におけるストレージ管理装置 2 およびネットワーク装置 3 の構成は、第二の実施形態と同様である。

【0147】

第六の実施形態における各装置間の通信シーケンスは、マスタストレージ装置 1 がホスト 16 に置き換わる点、リモートストレージ装置 5 がストレージ装置 15 に置き換わる点とを除き、第二の実施形態と同様である。

【0148】

以上で、第六の実施形態を説明した。第六の実施形態によると、ホスト 16 が、ストレージ装置 15 との間の複数のパスを用いて i S C S I プロトコルに基づいたデータの書き込みや読み出しを実行するシステムにおいて、あるパス上のネットワーク装置 3 で輻輳が発生した時、ストレージ管理装置 2 から割合変更要求を受けたホストが輻輳発生の結果送

信レートが低下したパスの割合を下げるができる。これにより、第一の実施形態と同様に、輻輳発生通知および輻輳回復通知を送信する機能を持たないネットワーク装置 3 を有するシステムでも、通信に係るスループット低下を最小限に抑えることができる。

【実施例 7】

【0149】

次に、第七の実施形態を第五の実施形態と相違する部分に限って説明する。第七の実施形態では、第三の実施形態と同様に、ネットワーク装置 3 が E C N 等を用いてホスト 16 に輻輳発生および輻輳回復を通知する。

【0150】

第七の実施形態におけるホスト 16 の構成は、主メモリ 601 に、設定受信プログラム 612、輻輳受信プログラム 613 および割合計算プログラム 614 が格納される点およびディスク装置 603 に、タスク管理テーブル 221 およびパス管理テーブル 222 が格納される点以外は、第五の実施形態と同様である。

【0151】

第七の実施形態におけるストレージ管理装置 2 の構成は、第三の実施形態と同様である。

【0152】

第七の実施形態では、第三の実施形態と同様に、システム管理者等がリモートコンピュータ登録画面 700 上で設定した各パラメータを、C P U 204 で実行される設定送信プログラム 215 がホスト 16 へ送信する。そしてホスト 16 の C P U 604 で実行される設定受信プログラム 612 が前記パラメータを受信し、その内容を各種テーブルのレコードに登録する。

【0153】

第七の実施形態における各装置間の通信シーケンスは、マスタストレージ装置 1 がホスト 16 に置き換わる点、リモートストレージ装置 5 がストレージ装置 15 に置き換わる点とを除き、第三の実施形態と同様である。

【0154】

第七の実施形態におけるホスト 16 のパス管理テーブル更新処理の動作手順は、第三の実施形態と同様である。

【0155】

以上で、本発明の第七の実施形態を説明した。第七の実施形態によると、ネットワーク装置 3 が、ホスト 16 へ E C N 等を用いて輻輳発生や輻輳回復を直接通知することにより、ストレージ管理装置 2 の機能が制限された装置を有するシステムでも、通信に係るスループット低下を最小限に抑えることができる。

【実施例 8】

【0156】

次に、第八の実施形態を第七の実施形態と相違する部分に限って説明する。第八の実施形態では、第四の実施形態と同様に、ホスト 16 自体が輻輳の発生や輻輳の回復を検出する。

【0157】

第八の実施形態では、第四の実施形態と同様に、ホスト 16 が、データの読み出し又は書き込みに使用する T C P コネクションにおいて A C K がタイムアウトした場合や、同じシーケンス番号を持つ A C K を 3 つ以上受信した場合に、輻輳発生とみなす。

【0158】

一方、輻輳発生からあらかじめシステム管理者等が設定した時間が経過した時あるいは T C P コネクションの輻輳ウィンドウサイズが、あらかじめシステム管理者等が設定したサイズを超過した時に輻輳回復とみなす。

【0159】

以上で、本発明の第八の実施形態を説明した。第八の実施形態によると、ホスト 16 が自律的に輻輳発生および輻輳回復を検出することにより、ネットワーク装置 3 に依存せず

通信に係るスループット低下を最小限に抑えることができる。

【図面の簡単な説明】

【0160】

【図1】 第一の実施形態におけるシステム構成例を示す図である。

【図2】 第一の実施形態におけるマスタストレージ装置およびストレージ管理装置の構成例を示す図である。

【図3】 第一の実施形態におけるネットワーク装置の構成例を示す図である。

【図4】 ネットワーク管理テーブルおよびタスク管理テーブルのデータ構造例を示す図である。

【図5】 パス管理テーブルのデータ構造例を示す図である。

【図6】 ネットワーク登録画面の表示例を示す図である。

【図7】 リモートコピータスク登録画面の表示例を示す図である。

【図8】 ネットワーク登録処理、パス登録処理およびリモートコピータスク登録処理の動作を示すフローチャートである。

【図9】 第一の実施形態における通信シーケンス例を示す図である。

【図10】 第一の実施形態におけるパス管理テーブル更新処理の動作を示すフローチャートである。

【図11】 第一の実施形態におけるパス管理テーブル更新処理の動作を示すフローチャートである。

【図12】 割合計算処理の動作を示すフローチャートである。

【図13】 第二の実施形態におけるストレージ管理装置およびネットワーク装置の構成例を示す図である。

【図14】 第二の実施形態における通信シーケンス例を示す図である。

【図15】 第三の実施形態におけるマスタストレージ装置およびストレージ管理装置の構成例を示す図である。

【図16】 第三の実施形態における通信シーケンス例を示す図である。

【図17】 第三の実施形態におけるパス管理テーブル更新処理の動作を示すフローチャートである。

【図18】 第五の実施形態におけるシステム構成例を示す図である。

【図19】 第五の実施形態におけるホストの構成例を示す図である。

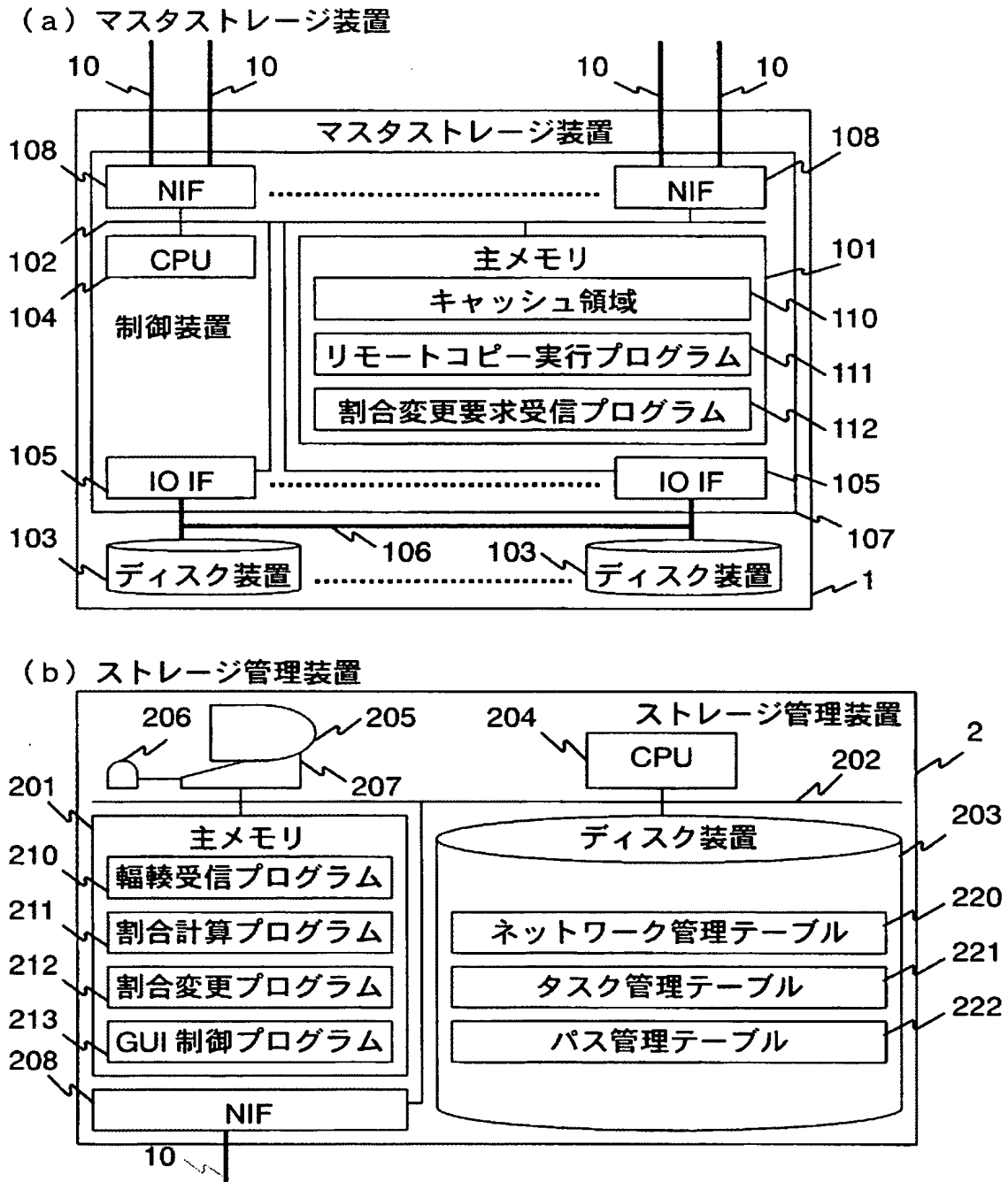
【符号の説明】

【0161】

1…マスタストレージ装置、2…ストレージ管理装置、3…ネットワーク装置、4…ホスト、5…リモートストレージ装置、6…端末、7…IP-SAN、8…WAN、9…LAN、10…通信線、11…マスタサイト、12…リモートサイト、13…管理用ネットワーク、14…通信線、15…ストレージ装置、16…ホスト。

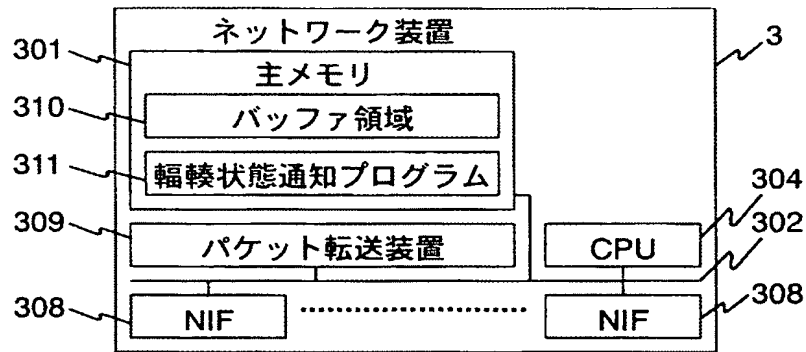
【図 2】

図 2



【図 3】

図 3



【図 4】

図 4

(a) ネットワーク管理テーブル 220

2201 ネットワーク ID	2202 装置 ID	2203 IP アドレス
WAN01	DEV01	12.34.56.78
WAN01	DEV02	12.34.56.80
...

(b) タスク管理テーブル 221

2211 タスク ID	2212 イニシエータ iSCSI 名	2213 イニシエータ LUN	2214 ターゲット iSCSI 名	2215 ターゲット LUN
TSK01	mcu0101	mcu0101	rcu0101	mcu0101
TSK02	mcu0102	mcu0102	rcu0102	mcu0102
...

2216 割合変更 フラグ	2217 割合 変更率
1	50
1	50
...	...

【図 5】

図 5

パス管理テーブル 222

パス ID	タスク ID	ネットワーク ID	デフォルト 割合	割合	輻輳回数
PATH01	TSK01	TSK01	50	25	1
PATH02	TSK01	TSK01	50	75	0
...

イニシエータ IP アドレス	ターゲット IP アドレス
192.168.0.1	192.168.2.1
192.168.1.1	192.168.3.1:3261
...	...

【図 6】

図 6

ネットワーク登録画面

ネットワーク ID:

ネットワーク装置

装置 ID:

輻輳発生・回復通知のソース IP アドレス:

追加

削除

装置リスト:

ネットワーク ID	装置 ID	ソース IP アドレス
WAN01	DEV01	12.34.56.78
WAN01	DEV02	12.34.56.79

OK

【図 7】

図 7

リモートコピータスク登録画面

タスク ID:

イニシエータ iSCSI 名:

イニシエータ LUN:

ターゲット iSCSI 名:

ターゲット LUN:

パスの設定

パス ID:

使用するネットワーク ID:

イニシエータ IP アドレス:

ターゲット IP アドレス:

割合:

パスリスト:

タスク ID	パス ID	イニシエータ IP アドレス	ターゲット IP アドレス	割合
TSK01	PATH01	192.168.0.1	192.168.2.1:3260	50
TSK01	PATH02	192.168.1.1	192.168.3.1:3260	50

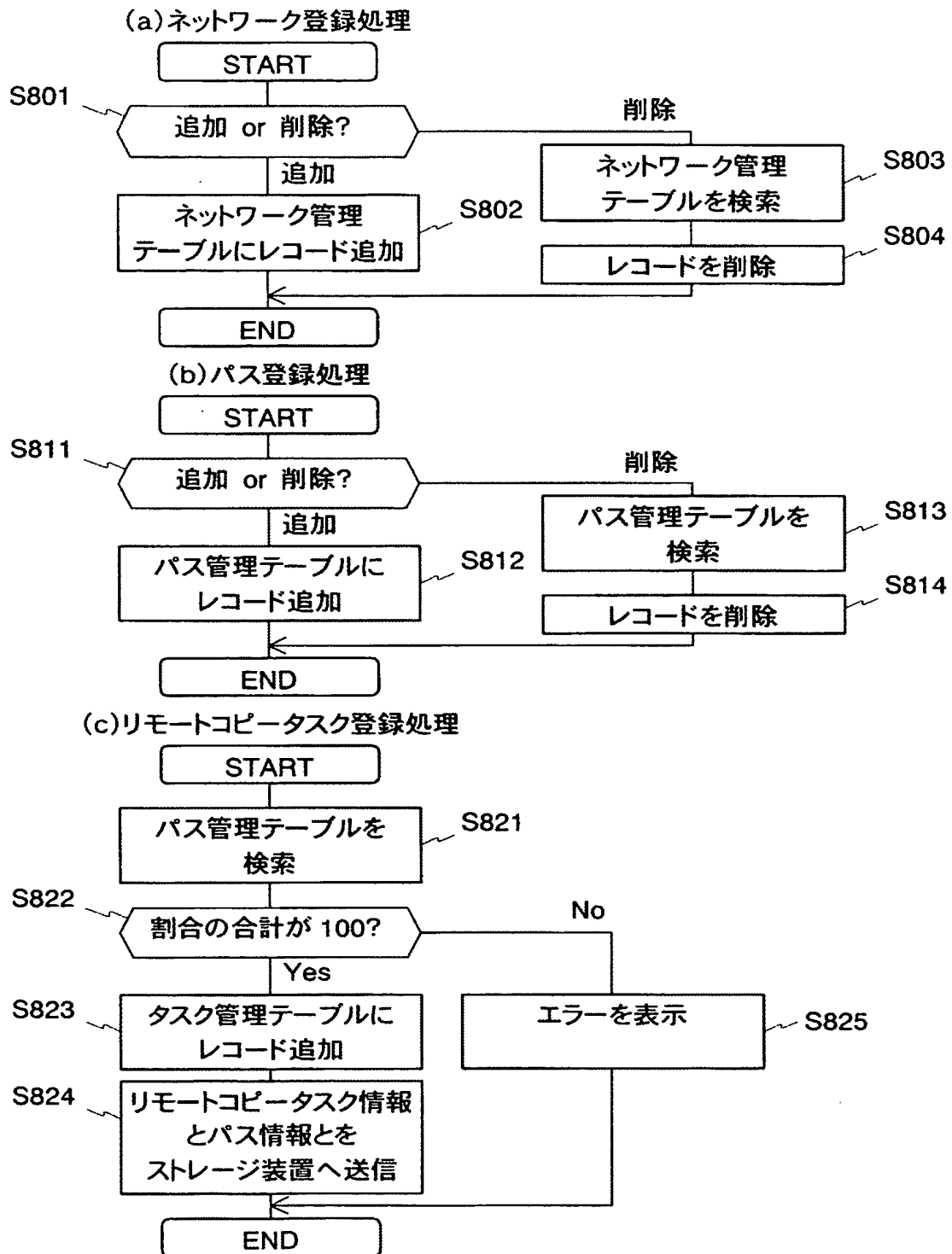
負荷分散の設定

☒ 輻輳時に負荷分散の割合を変更

輻輳したパスの割合変更率(%):

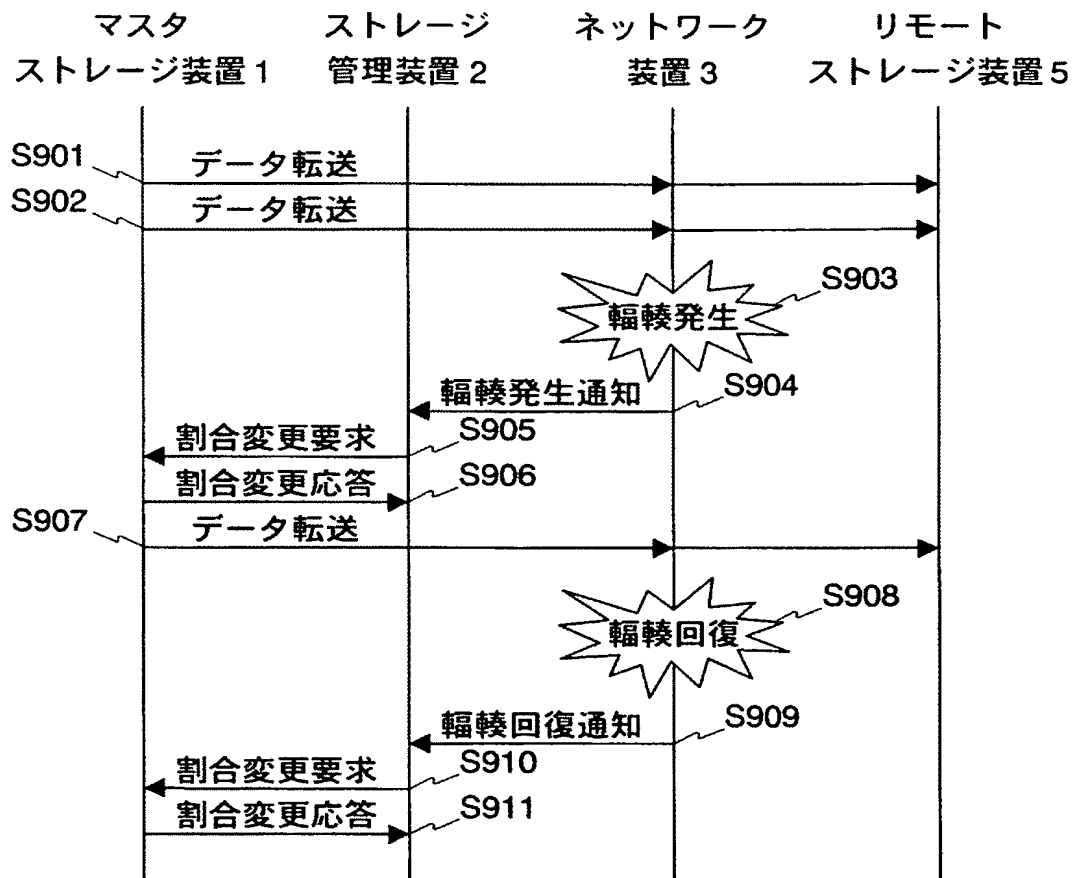
【図 8】

図 8



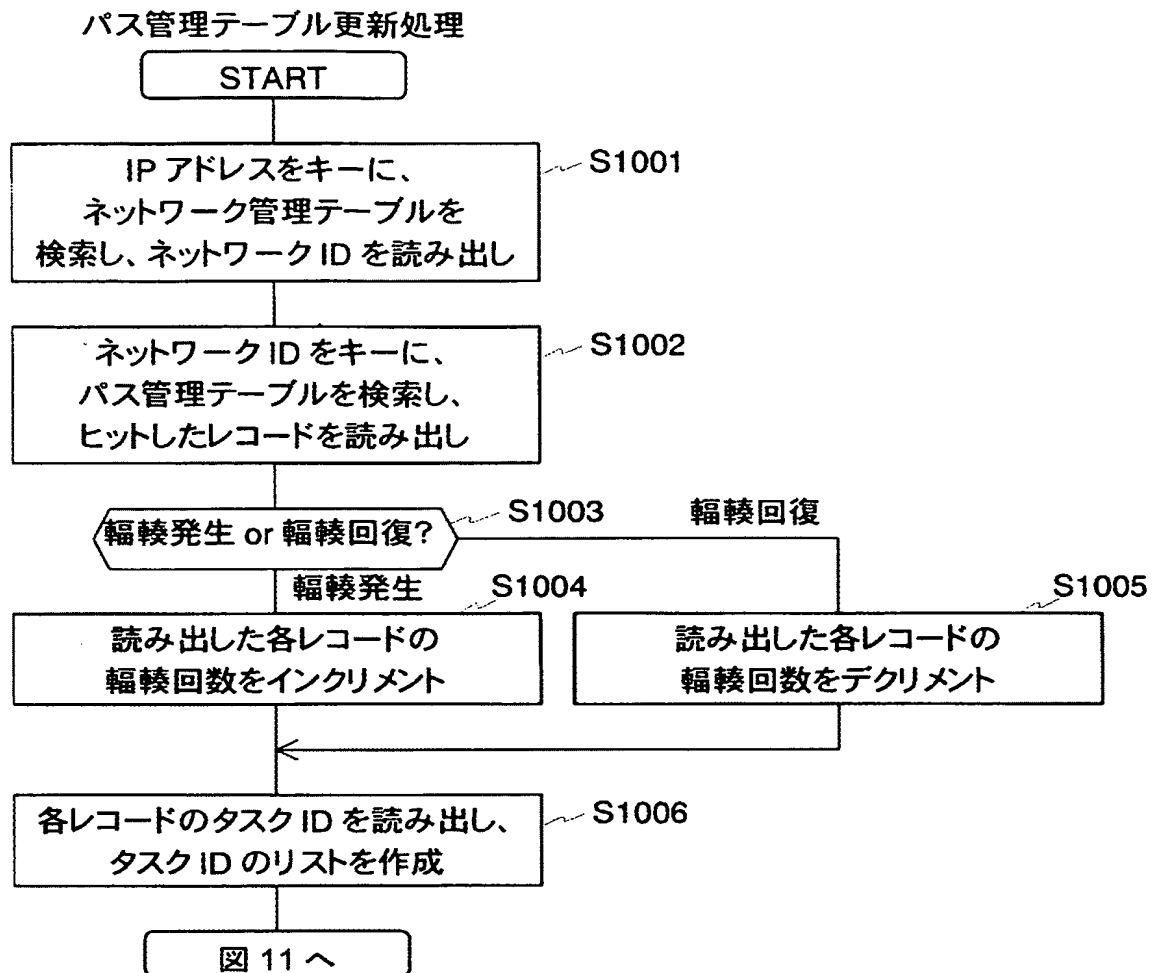
【図 9】

図 9



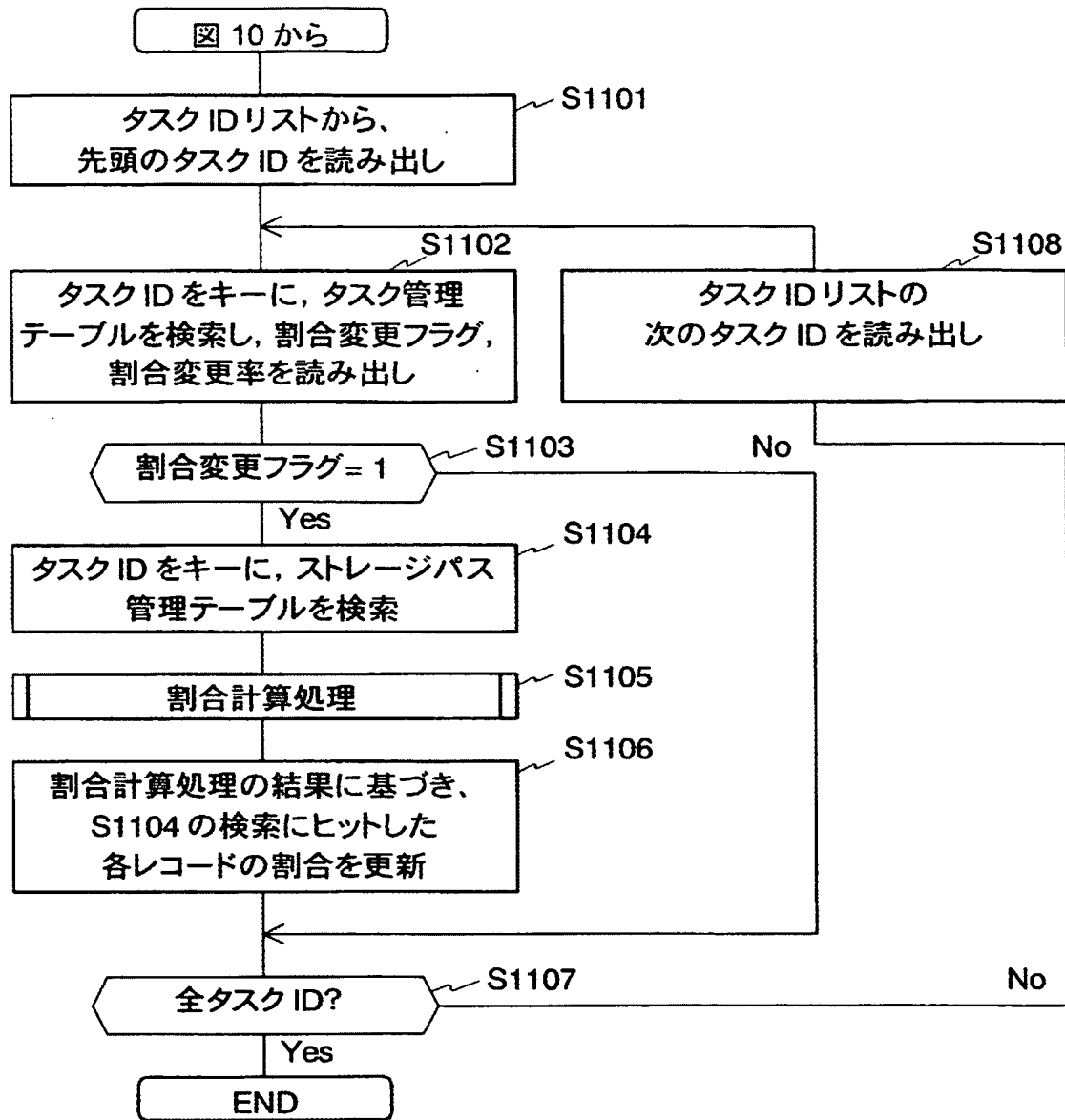
【図 10】

図 10



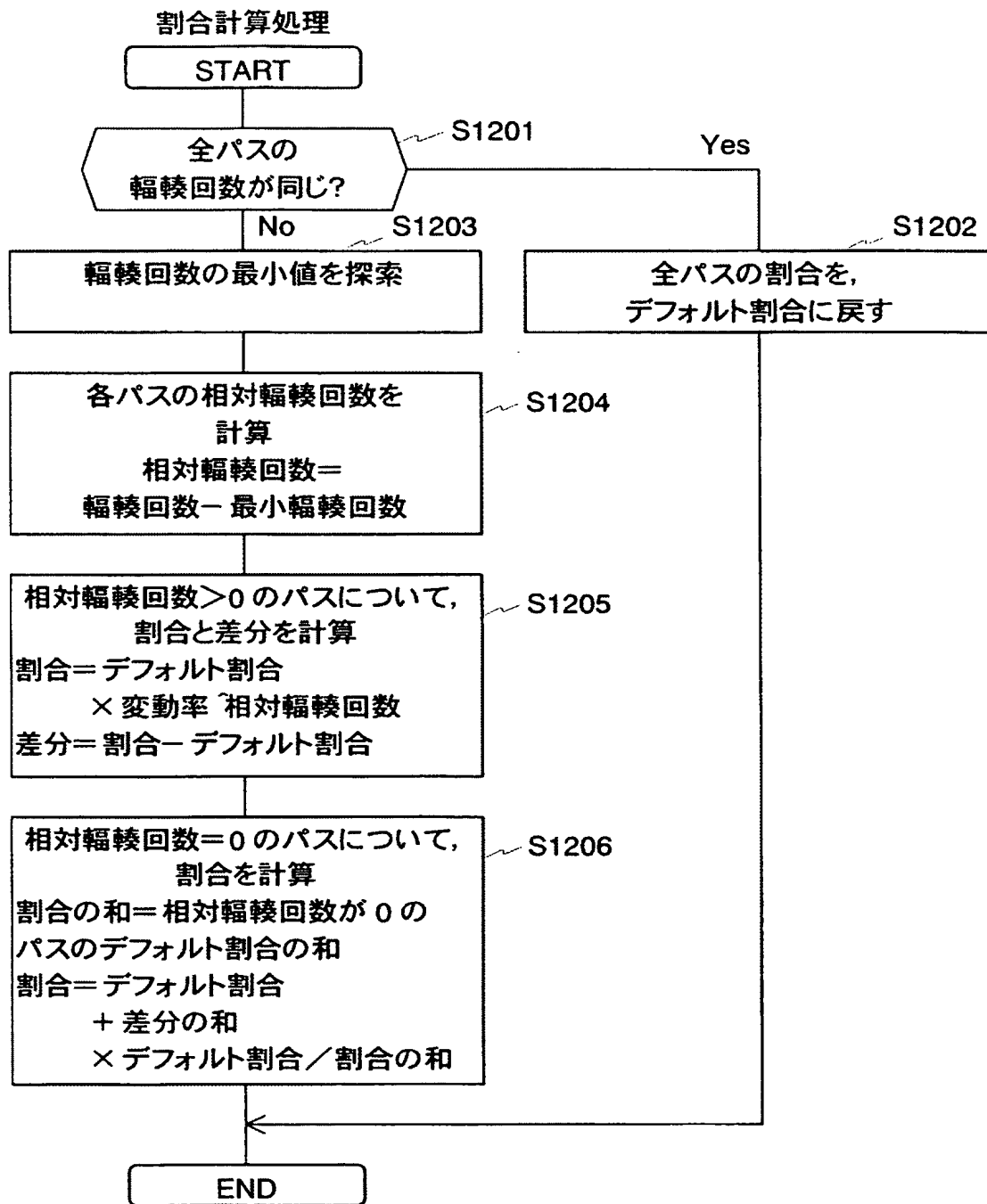
【図 11】

図 11



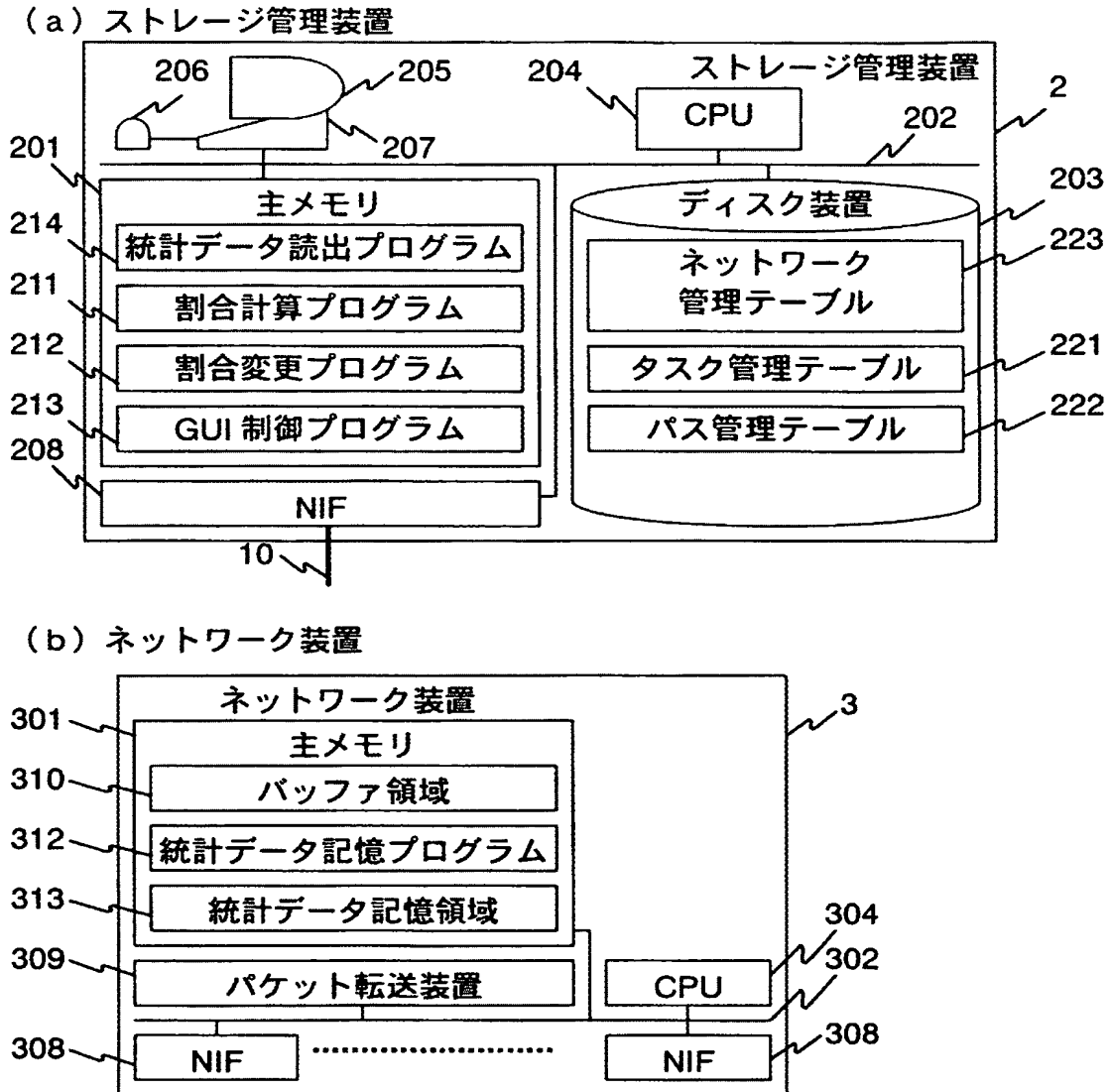
【図 12】

図 12



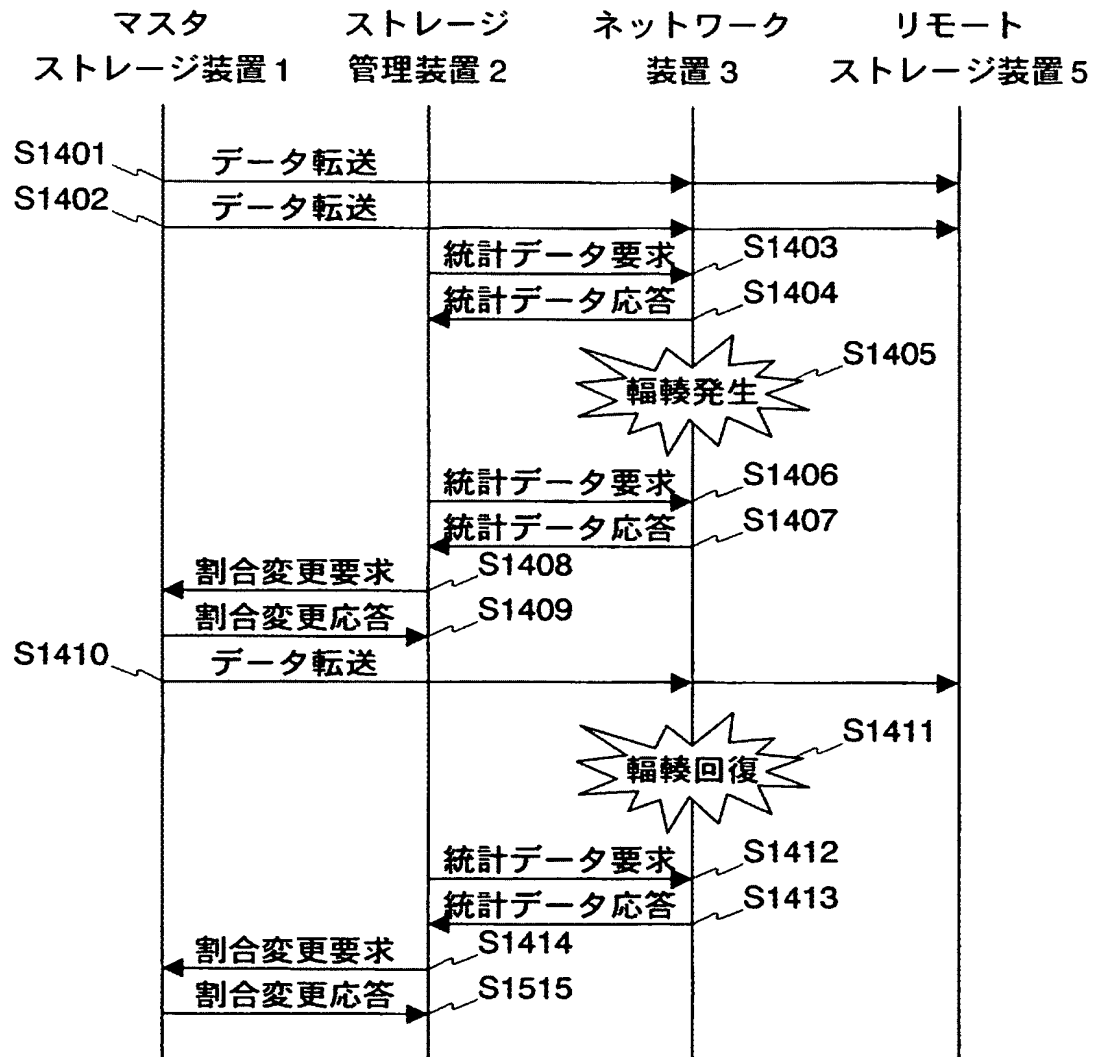
【図 13】

図 13



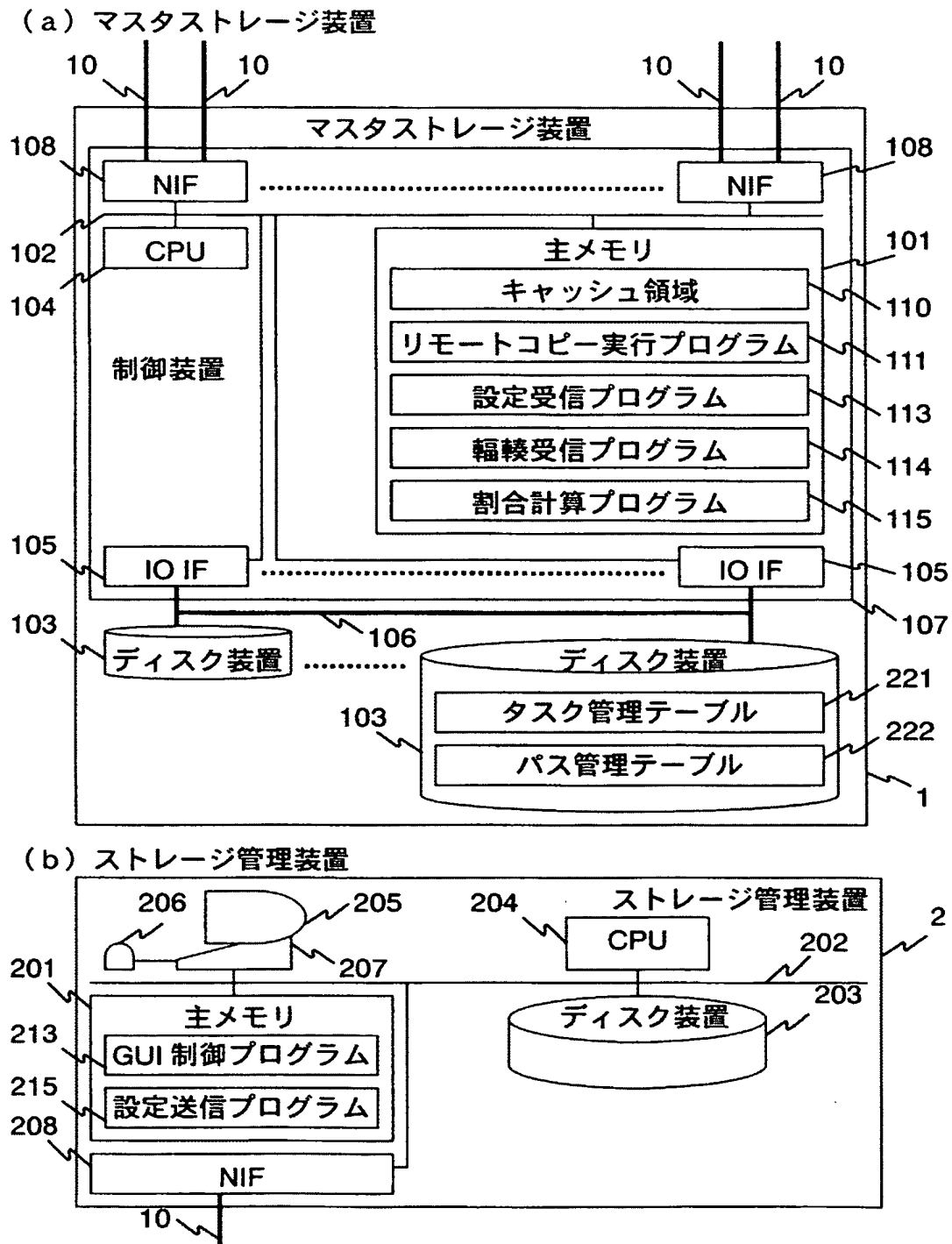
【図 1 4】

図 1 4



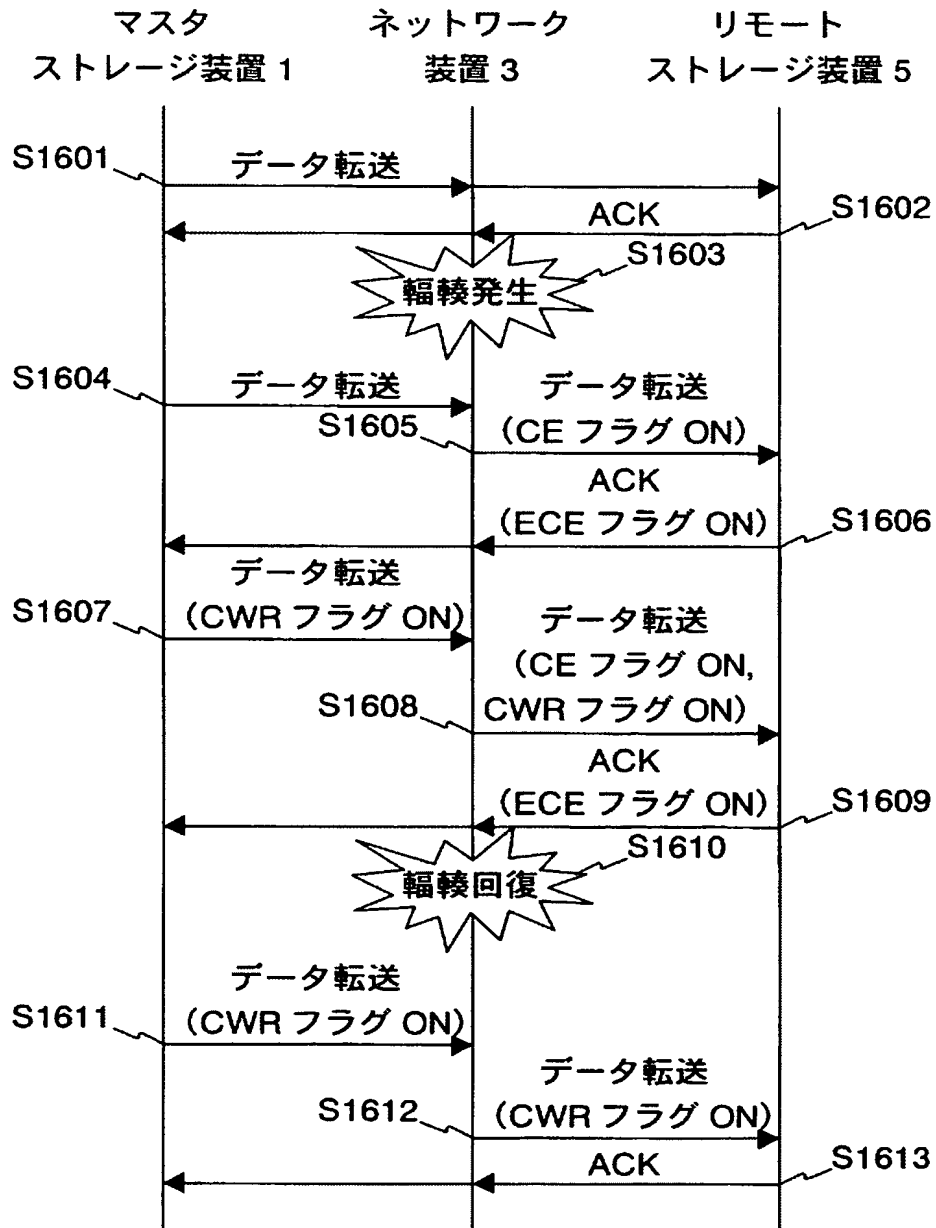
【図 15】

図 15



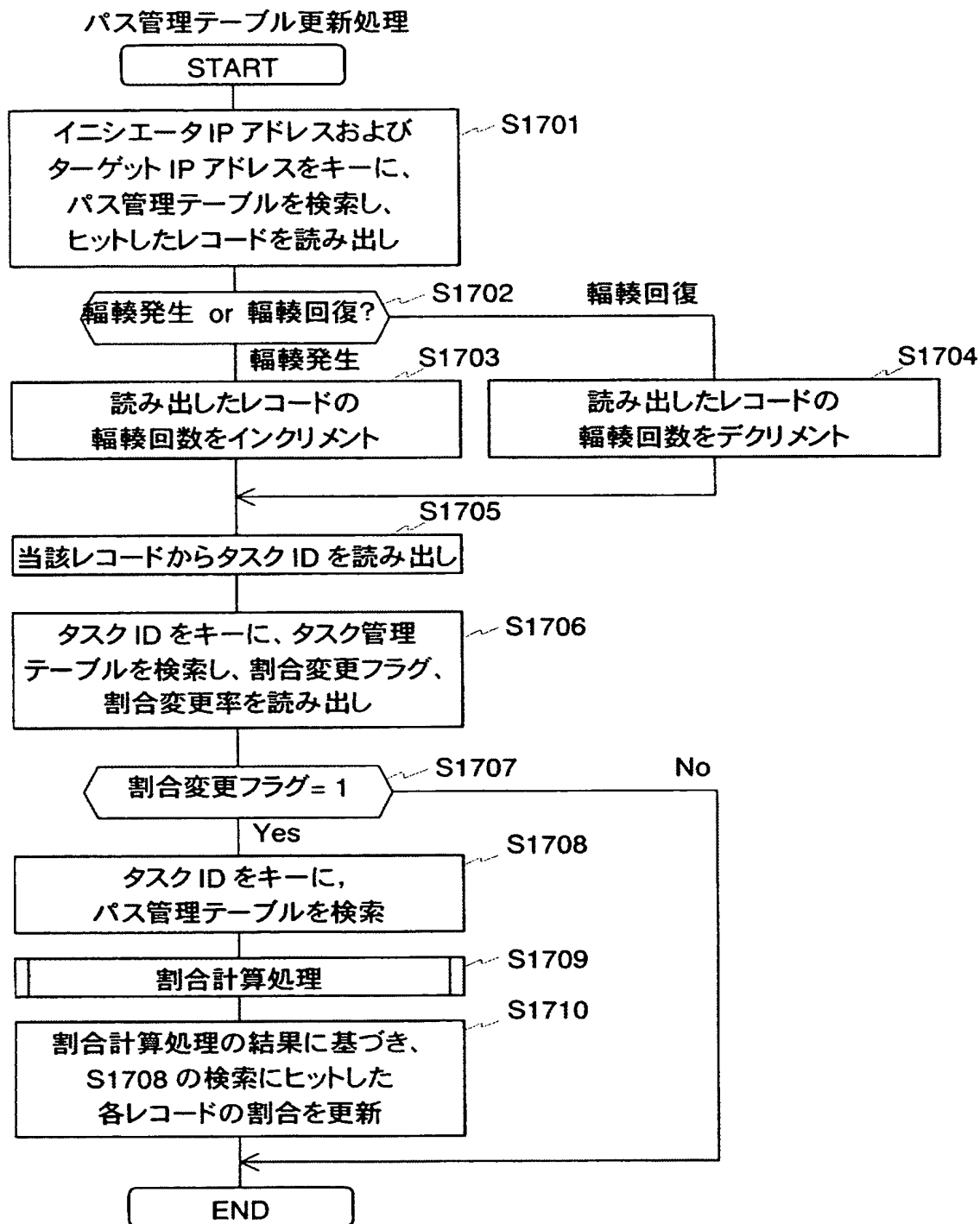
【図 16】

図 16



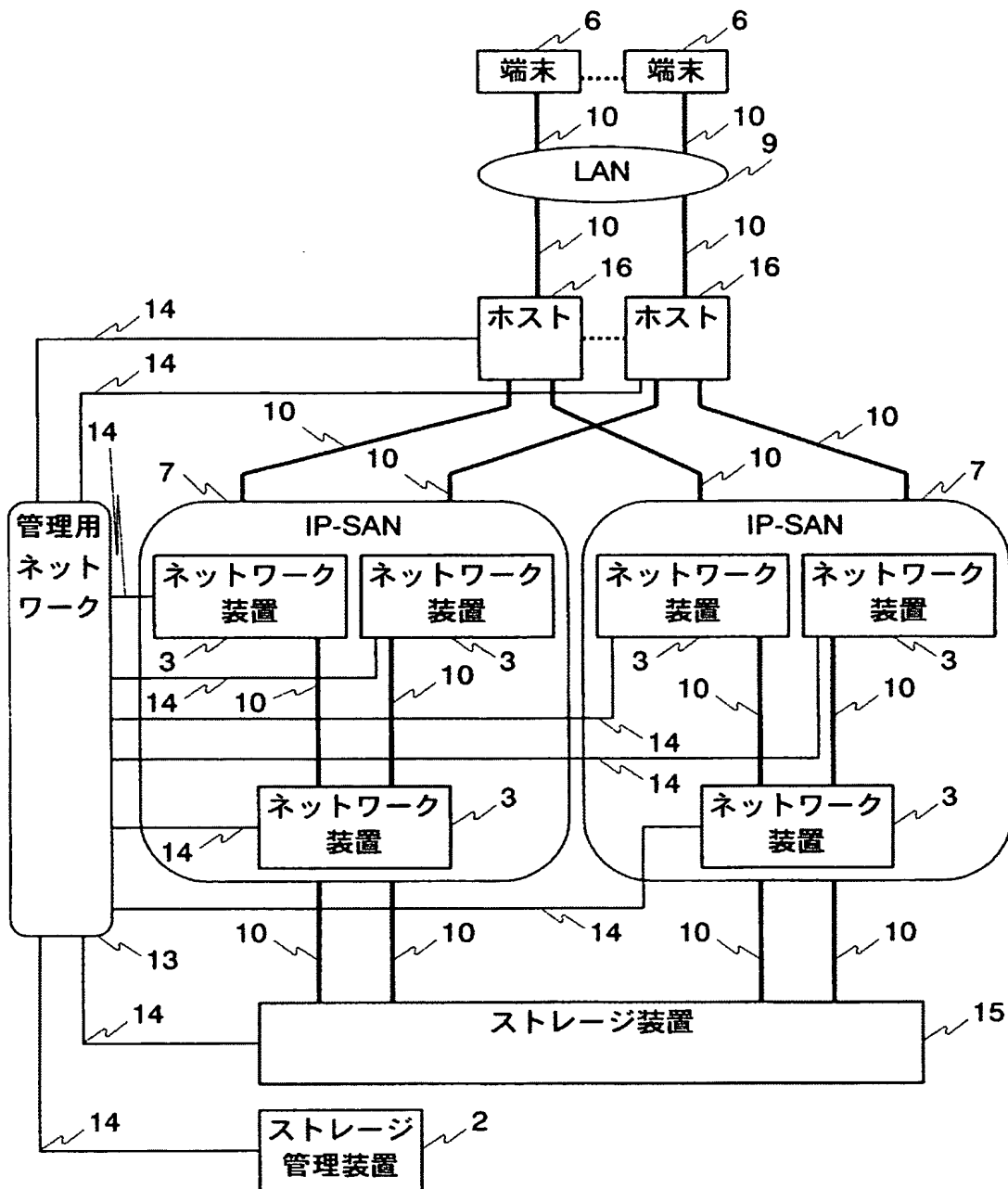
【図 17】

図 17



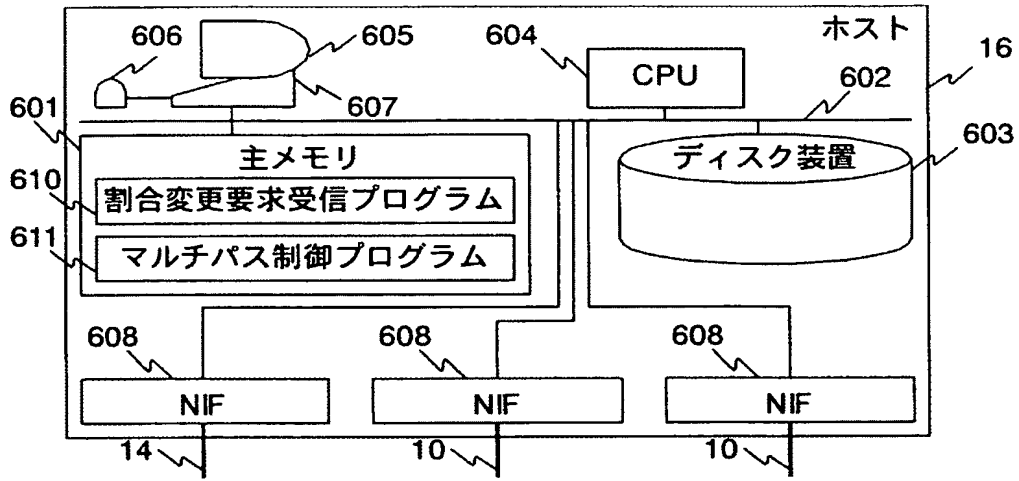
【図 18】

図 18



【図 19】

図 19



【書類名】 要約書**【要約】****【課題】**

計算機あるいはストレージ装置が、複数の経路を用いてストレージ装置へアクセスするシステムにおいて、経路上で輻輳が発生した時の、スループットの低下を防止する。

【解決手段】

計算機、ストレージ装置およびストレージ管理装置がネットワークで接続され、計算機あるいは第一のストレージ装置が、ネットワーク内の複数の経路を用いて、第二のストレージ装置へアクセスし、かつ計算機あるいは第一のストレージ装置が、あらかじめ設定された割合に基づき複数の経路間の負荷分散を行う計算機システムにおいて、経路上で輻輳が発生したことを検出し、当該経路の輻輳時の比重を計算し、輻輳時の割合と設定された割合との差分を、計算機あるいは第一のストレージ装置と第二のストレージ装置との間の、その他の経路の割合へ割り振る。

【選択図】 図 2

認定・付加情報

特許出願の番号	特願 2 0 0 4 - 0 2 2 4 6 3
受付番号	5 0 4 0 0 1 5 3 2 4 0
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 6 年 2 月 2 日

< 認定情報・付加情報 >

【提出日】 平成16年 1月30日

特願 2 0 0 4 - 0 2 2 4 6 3

出 願 人 履 歴 情 報

識別番号 [0 0 0 0 0 5 1 0 8]

1. 変更年月日	1 9 9 0 年 8 月 3 1 日
[変更理由]	新規登録
住 所	東京都千代田区神田駿河台 4 丁目 6 番地
氏 名	株式会社日立製作所